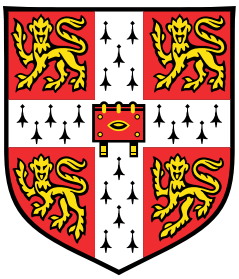
Quantifying the pro- and antimutagenic roles of DNA damage and repair



Nadezda Volkova

European Bioinformatics Institute
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Lucy Cavendish College
September 2019

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university.

This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation does not exceed the specified length limit of 60,000 words as defined by the Biology Degree Committee.

September 2019

Nadezda Volkova

Summary

Quantifying the pro- and antimutagenic roles of DNA damage and repair

Nadezda Volkova

Genome integrity is essential to the survival of any living organism. The genome is constantly challenged by a multitude of endogenous and exogenous mutagenic factors such as environmental exposures or replication errors. Therefore, evolution has supplied cells with a number of repair mechanisms to protect their genetic information; however, excessive exposures or defects in the repair machinery can lead to the accumulation of deleterious mutations which may cause a range of diseases including cancer.

Different mutational processes often leave behind characteristic patterns of mutations, so-called mutational signatures. Mutational signature analysis of tumours has gained a lot of attention recently, because it may reveal carcinogenic exposures and also therapeutic vulnerabilities. So far, over 50 mutational signatures have been identified using pattern recognition in large cancer cohorts, reflecting the action of a range of known mutagenic processes, such as UV light, tobacco smoke or mismatch repair deficiency, but for many mutational signatures an underlying generative process is still unknown. The search for the causes behind a given mutational signature is further complicated by the fact that every alteration in the DNA results from failed or incorrect repair of a DNA lesion, hence there are two factors which jointly shape the mutational spectrum of any mutagenic process.

In this thesis, I quantify the variability of mutational signatures in model organisms and in human cancer and explore the diversity of DNA damage-repair interactions. Using data from a large mutagenesis screen in *C. elegans*, including over 50 DNA repair deficient genetic backgrounds, 12 genotoxins and nearly 200 combinations thereof, I characterise the mutational spectra and genomic features of a range of DNA repair deficiencies, and describe the mutational signatures of genotoxins across multiple genetic backgrounds. Importantly, the mutagenic contributions of genetic and mutagenic factors can vary de-

pending on the DNA repair components available: over 35% of genotoxin-knockout combinations demonstrated a measurable effect on the mutation rate compared to expected values, and about 10% also presented a new mutational spectrum.

Analysis of mutational signatures in cancer exomes demonstrates the relevance of *C. elegans* results to cancer investigation. Mismatch repair deficiency patterns extracted from *C. elegans* are comparable to those in gastrointestinal tumours, and help to dissect convoluted mutational processes. The antagonism between DNA damage and repair drives variability in cancer genomes as well: the observed interaction effects were low in magnitude, but evolutionary considerations suggest that cancer risk may be substantially elevated even by small increases in mutagenicity.

In summary, this thesis presents the first comprehensive analysis of mutagenic DNA damage-repair interactions using experimental and cancer data. The results show that mutations result from the opposing pro- and anti-mutagenic forces of DNA damage and repair, which shape mutational signatures in highly variable ways. This variation has to be acknowledged and integrated into mutational signature analysis to ensure reliable interpretation and applicability in clinical oncology. Lastly, the cross-species comparison shows that the fundamental laws of mutagenesis are acting similarly across eukaryotic organisms reminding that many mutational processes fuelling tumorigenesis are not exclusive to cancer, but also drive variation and the evolution of species.

Acknowledgements

In the first place, I would like to thank my amazing supervisor Moritz for his support, patience, enthusiasm and passion for truth. His guidance and criticism shaped my development as a scientist and widened the horizons of my understanding of science. Without a doubt, I was very fortunate to be able to watch, talk to and learn from such a brilliant researcher.

I am also very grateful to all the former and current members of the Gerstung group, in particular to Sara Killcoyne, Santi Gonzalez, Harald Voehringer, Yu Fu, Rui Costa, Jose Almeida and Artem Lomakin – you all made my PhD full of exciting conversations, support and fun. Being a part of this awesome group was a great time, and also shaped my communicative and critical skills to the state where I am now. Another wave of appreciation has to be given to my fantastic PhD mates, especially to Marta Julia Strumillo, who is a true star and without whom I would not make it till the end in my right mind, and also to students from other institutes – Vladislava, German, Andrey – who helped me more than they think.

Next, I owe a huge thank you to my collaborators, Bettina Meier and Anton Gartner, who provided me with all the necessary help and guidance to be able to dive into the world of mutagenesis and taught me how to communicate between the worlds of experimental biology and statistics.

I would like to thank Lucy Cavendish college for being a welcoming and extremely helpful environment, and all the wonderful ladies whom I met over the years of singing in the Cavendish Chorale for being such a vibrant and broad-minded community.

A special thanks goes to my family, whose love and support go far beyond any geographical borders, and who raised me in the spirit of curiosity and taught me to work hard, love science and never give up. The last huge thank you is for Vlad - for his love and patience, for believing in me more than I did, and for always being there for me.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Every system which stores and transmits information is prone to compression loss and transmission errors. According to Shannon's noisy-channel coding theorem, there is a maximum limit of error-free transmission rate for every channel with a certain noise rate, which makes it impossible to send over a signal completely error-free in presence of noise. In this regard, a genome of a cell is no different: DNA has to be copied every time a cell divides, which exposes the valuable genetic information to damage induced by the environment and to errors arising via replication. The task of correctly propagating genetic information is crucial to the survival of a species, yet it is the inevitability of random errors that drives the evolution. It is the balance between the two that allows for the development of such huge diversity of life forms capable of adapting to changing environments.

Moreover, DNA within a living cell experiences pressures other than replication. Functional regions are constantly accessed by various enzymes conducting transcription or regulation thereof. Various exogenous mutagens are capable of altering the chemical or physical structure of DNA, even when it is tightly compressed. For gamete production, DNA may have to undergo crossover, mixing the information between the two copies with a risk of loss or alteration of genetic information. The chemical environment of the cell also poses a threat to genome stability via spontaneous hydrolysis reactions and the presence of reactive oxygen species which can modify DNA bases. The chance of acquiring damage is low for each individual nucleotide, but becomes high when considered on the scale of 12 billion bases of a diploid genome. To cope with a multitude of damage-inducing processes, cells have developed a rich toolkit of DNA repair mechanisms which aim to restore the DNA to the original state when possible, or at least minimise the consequences of DNA alterations.

There is a long history of understanding how different traits can be inherited and modified. The first principles of evolution and heredity were laid as a foundation of

genetics in the 19th century with Charles Darwin's seminal manuscript on the origins of species and the works of Gregor Mendel, which were rediscovered by Hugo de Vries at the end of 19th century. The concept of mutation and its effects on the organism were further enhanced by the brilliant experiments conducted by the lab of Thomas Morgan in the early 20th century.

The discovery of mutagenic abilities of X-rays in the 1920s opened the door for mutagenesis experiments and led to the formulation of "target theory", a unified theory of mutation. In the 1930s, with the discovery of chemical mutagens and the introduction of micro-organisms as a model system, experimental studies of mutations reached a new level allowing for the characterisation of mutagenic properties of different substances. The DNA structure discovered by Watson and Crick in 1953 gave the mutagenesis field biochemical and biophysical perspectives.

The introduction of sequencing in the 1970-s allowed scientists to look at mutations at the single-base resolution. Thousands of studies were conducted, identifying mutations in individual genes and characterising their types and origins. Finally, the late 20th to early 21st century's era of high-throughput sequencing flooded genetics with an enormous amount of data on mutations in model organisms, human germ line (such as the 1000 Genomes project and UK Biobank), as well as cancer (including TCGA, ICGC, MSK-IMPACT).

Mutation acquisition is one of the most basic biological processes underlying adaptation, immunity, and disease, but despite of this, many aspects of it are still poorly understood. In the Section 1.1 I will provide an extensive overview of known sources and types of DNA damage as well as DNA repair mechanisms. The mutations resulting from the tussle of DNA damage and repair do not just affect individual genes or genome regions, but also generate patterns which can reveal the mechanistic principles of mutagenesis. A review of the recognition and utility of such patterns will be presented in Section 1.2. Lastly, in Section 1.3 I will point out several underexplored aspects of mutational signatures which I will address in this thesis work.

1.1 DNA damage and DNA repair

A common feature of any living organism is the ability to pass on their genetic information in the form of DNA. The fidelity of DNA replication may be challenged by a multitude of endogenous and exogenous damaging factors. To protect the genome from deleterious mutations, cells are equipped with a range of DNA repair mechanisms to preserve the DNA in its original unaltered state, or at least to repair major lesions that would make replication impossible at a cost of less damaging mutations.

Mutations are alterations in the DNA sequence, and they result from unrepaired (or incorrectly repaired) lesions: damage to the chemical or physical structure of the double helix or its components conferred by spontaneous reactions, polymerase errors, exogenous agents or endogenous mutagenic processes. Such mutations can be measured on the genome-wide scale using next-generation sequencing (NGS), which allows detecting changes at a single-base resolution in the DNA sequence by comparing it to the reference. In this section, I will describe the most common types and sources of DNA damage and the mutations they cause, as well as give an overview of the pathways which conduct DNA damage sensing and repair. For clarity, the term DNA damage in this chapter will only refer to the primary damage (lesions and nucleotide modifications), whereas in the chapters to follow, this term will denote both the primary damage and the mutations caused by a particular mutational process.

1.1.1 Types of DNA damage

Damage to DNA is occurring constantly: the nuclear DNA of an average cell of the human body experiences over 70,000 damage episodes per day, most of which are being efficiently repaired, as the resulting mutation rate is as low as 10^{-10} mutations per base per cell division, corresponding to less than one mutation per genome (Bernstein et al. 2013). Apart from being a source of important changes for evolution and adaptation, DNA damage is involved in many processes less beneficial to an individual organism: it is implicated in ageing (Hoeijmakers 2009), and it is the main component of neoplastic development (Stratton, Campbell, and Futreal 2009).

Despite a high number of damage instances occurring in a cell over a single cell cycle, most of this damage in a healthy cell will be correctly repaired (Sancar et al. 2004). Only the lesions that manage to persist through the replication, or large-scale events with a higher chance of incorrect restoration (of DNA), will result in a somatic mutation in one or both of the daughter cells. If the damage is excessive and cannot be repaired within a reasonable time frame, it can trigger cell cycle arrest in proliferating cells and put them in senescence, forcing them to stop dividing until the damage is repaired. If the damage still cannot be repaired, damaged cells will undergo programmed cell death – apoptosis (Hoeijmakers 2009). The estimates of the steady-state number of DNA lesions, i.e. when there is a balance between lesion formation and repair, report about 30,000 abasic sites being present in an average cell of a rat. However, this number may vary substantially across different tissues (Swenberg et al. 2010).

DNA damage types range from chemical modifications affecting an individual DNA base to the changes in the spatial structure of DNA (Ward 1994). These kinds of damage

Sources of DNA damage	Type of damage
Spontaneous hydrolysis	Apurunic / apyrimidinic sites
Methylation of DNA bases	5-methylcytosine
Deamination of DNA bases	
Cytosine	Uracil
5-meC	Thymine
5-hydroxymeC	5-hydroxymeU
Adenine	Hypoxanthine
Guanine	Xanthine
ROS/RNS	8-oxoguanine, abasic sites, SSBs, deamination
Replicative errors	Mismatches
Alkylating agents	Base alkylation
PAHs	Bulky adducts
Intercalating agents	DNA structure distortion: Double- and single-strand breaks
Crosslinking agents	Intra- and interstrand crosslinks, DNA-protein crosslinks, monoadducts
UV-light	
Direct damage	CPDs, 6-4PPs
Indirect damage	Oxidative damage
Ionising radiation	
Direct damage	Double- and single-strand breaks
Indirect damage	Oxidative damage
Transposons	Translocations, insertions
Translesion synthesis	Mismatches (mostly with adenines)

Table 1.1: Types of DNA damaging agents and the damage they incur.

can be induced by a wide range of agents, some of which are listed in Table 1.1. Some of this damage can only be induced by external genotoxic agents, but there are also many internal processes that can cause damage to the DNA.

1.1.2 Endogenous sources of DNA damage

Many normal cellular processes can generate DNA damage by promoting chemical reactions between cell metabolites and DNA, or by directly producing the chemicals capable of altering the DNA structure or conformation (Tomasetti, Vogelstein, and Parmigiani 2013).

Spontaneous reactions

The structure of DNA relies on a huge amount of chemical bonds with varying stability (Lindahl 1993). Hydrolysis is a reaction cleaving the N-glycosidic bond of a nucleic acid

base. This happens spontaneously at a rate of 10,000 events per human cell per day, and it is mostly depurination as purines are more susceptible to loss due to hydrolysis (Lindahl and Barnes 2000). Apurinic or more generally abasic sites are mutagenic because the gap in the DNA strand must be filled upon replication, and the enzymes performing this reaction have low accuracy (Sale, Lehmann, and Woodgate 2012).

In addition, DNA bases and their modifications can spontaneously lose an amine group. Spontaneous deamination of cytosine turns it into uracil, which happens about 70-200 times per human cell per day (Lindahl 1993). Cells also possess a special system of enzymes deaminating cytosines in single-stranded DNA as a part of viral protection or somatic hypermutation mechanisms (Knisbacher, Gerber, and Levanon 2016), namely AID and APOBEC. As cytosines in ssDNA are turned into uracils, they have to be excised by special enzymes ensuring there is no uracil in the DNA. Depending on whether a uracil or an abasic site is left in the DNA upon the next replication, it can lead to C>T or C>G mutations, respectively (Taylor et al. 2013).

Deamination of adenine happens less often and leads to hypoxanthine, a base modification that can pair with cytosine (Karran and Lindahl 1980). Guanines can also be spontaneously deaminated producing xanthine - a base that typically pairs with cytosine and is not mutagenic (Lindahl 1993).

In mammals, an additional source of mutations is DNA methylation. 5-methylcytosine is prone to spontaneous deamination which results in thymine, a DNA base that cannot be detected by glycosylases and creates a mismatch mostly detectable by the mismatch repair system. Off-target methylation of 3-position in cytosine can also be mutagenic (Delaney and Essigmann 2004).

Replication errors

Replication is one of the most essential and complex processes in the cell, which can also generate mutations as a result of nucleotide misincorporation or polymerase slippage. Stalling of replication fork can lead to double-strand breaks (DSBs) that can be lethal to the cell (Jackson 2002), while replication-transcription collisions create deletions or duplications and base substitutions at promoters of the genes (Sankar et al. 2016).

Typically, replication starts at special positions in the genome termed origins of replication, 30,000 to 50,000 of which are activated in a human cell during cell division (Prioleau and MacAlpine 2016). Helicases open the double-stranded DNA, forming a replication fork. Then, three kinds of replicative polymerases are recruited: polymerase α, which serves as replicative primase, and leading strand polymerase ε and lagging strand polymerase δ.

Three mechanisms are employed by the cells to ensure the fidelity of replication.

Firstly, the replicative polymerases typically have high selectivity towards the right nucleotide (Kunkel and Bebenek 2000). If they do insert a wrong one, these polymerases have an exonuclease domain that allows detecting and removing a mismatch straight away. However, this system is not perfect and can miss about one mismatch per 10 Mbps (Kunkel and Bebenek 2000), and cannot deal with situations where the repetitive sequence context causes the newly synthesised and template strand to bind at the wrong repeat resulting in a set of unpaired bases in one strand. Hence, there is a third mechanism called mismatch repair that screens the DNA for such errors right after replication and repairs them in the newly synthesised DNA.

The availability of the substrate for polymerases can also have an impact on the error rate. Total depletion of nucleotide (dNTP) pool can lead to replication failure and cell death (Laureti et al. 2013). It has been reported that the mutations attributable to polymerase errors are enriched in late replicating regions, which may be also connected to the decreased availability of dNTPs (Koren et al. 2012).

Moreover, DNA bases can exist in several isomers which are chemically identical but can potentially pair with wrong bases. The normal DNA usually contains the keto form of bases, but it can turn into enol and imino forms upon a tautomeric shift. These events, however, are rare under normal physiological conditions (Abou-Zied, Jimenez, and Romesberg 2001).

Prolonged exposure of single-stranded DNA to the cell environments can also cause damage to the DNA. Repetitive regions can form various secondary structures while left single-stranded during replication leading to replication stalling and deletions or insertions upon replication restart (Huang et al. 2017b). Similarly, overexpression of APOBEC enzymes can lead to APOBEC-mediated cytosine deamination during replication, generating a pattern of mutations often found in cancers (Roberts et al. 2013).

Oxidative damage

Chemically reactive substances containing oxygen such as peroxides, superoxides, hydroxyl radicals, singlet oxygen, or alpha-oxygen can trigger oxidative damage to the DNA. Reactive oxygen species (ROS) are metabolites stemming from multiple normal chemical reactions in the cell and participate in cell signalling and homeostasis. Typically, they are produced as a by product of oxidative phosphorylation in mitochondria, but can arise also in response to environmental stress such as ionising radiation, heavy metals and pollutants (Devasagayam et al. 2004).

ROS can induce base modifications such as 8-oxoG, abasic sites and DNA breaks, and take part in the creation of aldehydes that can crosslink DNA. In total, they can generation over 20 types of DNA lesions (De Bont and Van Larebeke 2004). On average,

natural oxidative damage causes about 12,000 lesions per human cell per day (De Bont and Van Larebeke 2004).

Mobile genetic elements

Transposons and other mobile genetic elements can move across the genome potentially disrupting the structure of a gene or functional element. The relocation of these elements causes DNA damage response, which can, in turn, recruit an error-prone repair pathway to repair the damage. Mobile genetic elements can cause gene duplications and fusions (Izsvák, Wang, and Ivics 2009).

1.1.3 Exogenous agents incurring DNA damage

Various physical, chemical or biological agents can induce DNA damage, either directly or indirectly. Pathogenicity of exposure can stem from the damage to the cell and its functions (cytotoxicity), or from damage caused to the DNA (genotoxicity). Many of the substances listed as carcinogens are genotoxins causing mutagenic alterations of the DNA (Health and Services 2016). I describe the most common exogenous genotoxins implied in human cancer development below.

Alkylating agents

Alkylating agents are chemicals capable of transferring an alkyl group to a DNA nucleotide. These agents may be monofunctional if they have only one binding site, or bifunctional if they can bind to two bases (or a base and a different molecule) (Watson 1964). Some alkylating agents such as quinones or ethylene oxide can occur naturally in the environment (Moore and Czerniak 1981, Garry et al. 1979), but most are chemically produced compounds often used as chemotherapy drugs (e.g. temozolomide, cyclophosphamide, mustard gas derivatives) due to their genotoxic capabilities (Kondo et al. 2010).

Monofunctional alkylating agents add an alkyl group to DNA bases causing base modifications that can result in base substitutions, single-strand breaks (SSBs), or stalled replication fork. In contrast, bifunctional alkylating agents create DNA or DNA-protein crosslinks and monoadducts leading to DNA breaks and replication fork stalling (Kondo et al. 2010).

Platinum-based cross-linking agents

Another class of drugs commonly used in cancer treatment are platinum-based agents such as cisplatin, oxaliplatin or carboplatin (Rabik and Dolan 2007). Due to the structure of their molecules, they can bind to more than one DNA base creating intra- and

interstrand crosslinks. Similar to bifunctional alkylating agents, they can also form bulky adducts and DNA-protein complexes. They mostly act on the adjacent N-7 position of guanine (Cohen and Lippard 2001). These chemicals are also termed 'alkylation-like' due to the similarity between their mechanism of action and that of bifunctional alkylating agents. If crosslinks remain unrepaired, they stall replication fork and cause single- and double-strand breaks.

Intercalating agents

Intercalating agents is a class of molecules that can fit themselves in between the DNA base pairs. Some of these agents are small enough to cause just a frameshift mutation during replication, and others can lead to transcription and replication blockage (Wakelin 1986). Alternatively, intercalating agents can target topoisomerase II (Nitiss 2009). During DNA replication or transcription, DNA helicases unwind the DNA to provide access to DNA or RNA polymerases. This intensifies the torsion of the double helix creating tangles and supercoils (Liu and Wang 1987, Vogelstein, Pardoll, and Coffey 1980). These structures are typically resolved by type II topoisomerase that cuts both DNA strands, unwinds the torsion and re-ligates the ends. Intercalating agents can interfere with this breakage-reunion process causing topoisomerase-mediated rearrangements (Nitiss 2009).

Polycyclic aromatic hydrocarbons

Polycyclic aromatic hydrocarbons (PAHs) are a type of chemicals often found in food, pollutants or fuel derivatives. They represent fused aromatic rings and are typically produced by the burning process. Many of them are highly carcinogenic, particularly so the benzo-[a]-pyrene, a component of tobacco smoke, and a food contaminant aflatoxin (Boffetta, Jourenkova, and Gustavsson 1997, Kew 2013).

In the process of being metabolised within the cell, PAHs turn into active molecules that can bind to DNA bases and create bulky adducts deforming the double helix and blocking transcription and replication (Xue and Warshawsky 2005).

Ionising radiation

Ionising radiation (IR) is a class of high-energy electromagnetic waves that can destabilise the structure of atoms in the human body, as well as alter otherwise inert molecules within the cell, turning them into reactive species (Lomax, Folkes, and O'Neill 2013).

Exposure to IR occurs during diagnostic radiology, cancer radiotherapy, and also as a consequence of environmental, occupational or accidental irradiation. IR has been long ago showed to be genotoxic and carcinogenic (Little 1993). Currently, radiotherapy is

one of the most effective cancer treatments applied to nearly 40% of UK cancer patients (Mayles 2010).

The main effect of IR on the cells is the DNA damage, mostly comprised of DSBs, which triggers senescence, apoptosis or other forms of cell death (Lomax, Folkes, and O'Neill 2013). IR can break chemical bonds, including those in DNA's and sugar phosphate backbone, leading to single- and double-strand breaks. In addition to direct harm, IR exposure modifies water and other organic molecules, turning them into reactive oxygen species and inducing oxidative damage (Lomax, Folkes, and O'Neill 2013).

Ultraviolet light exposure

Sun exposure is one of the main factors of skin cancer development. Sunlight represents UV-light irradiation composed of waves with different wavelengths. 95% of it consists of UV-A, the waves with lowest energy. Another 5% is UV-B, waves of intermediate wavelength. The short waves carrying the most energy, UV-C, are typically filtered by the ozone layer, but can be dangerous in areas with ozone layer anomalies (Herman et al. 1996).

UV exposure can cause both direct and indirect DNA damage. Direct damage comes in the form of cyclobutane pyrimidine dimers (CPDs) mostly affecting thymines and 6,4-photoproducts (6,4-PP) involving a cytosine and a thymine (Ikehata and Ono 2011). 6,4-PPs are more mutagenic than CPDs but occur at a third of the rate. Both of these lesions are typically repaired by the nucleotide excision repair, otherwise they stall replication fork and error-prone translesion synthesis has to be recruited (Sale 2013). Exposure to UV light can also cause indirect, or 'dark' damage: it has the right energy to excite electrons in molecules, leading to the formation of reactive chemical species and increasing the degree of oxidative stress (Ikehata and Ono 2011).

Biological agents

Viruses and bacteria can also induce DNA damage and are implicated in cancer development. Microbes can directly produce genotoxins (Žgur-Bertok 2013, Grasso and Frisan 2015) or cause prolonged inflammation that can result in elevated levels of oxidative stress (Kalisperati et al. 2017). Viruses can integrate their DNA into the host genome, disrupting the host's genes or triggering the production of viral enzymes damaging the host's DNA (Luftig 2014). About 12% of cancers worldwide are thought to be caused by oncoviruses (Schiller and Lowy 2010), most notably by Epstein-Barr virus, hepatitis B, several types of human papillomavirus, and mastadenovirus (Zapatka et al. 2018).

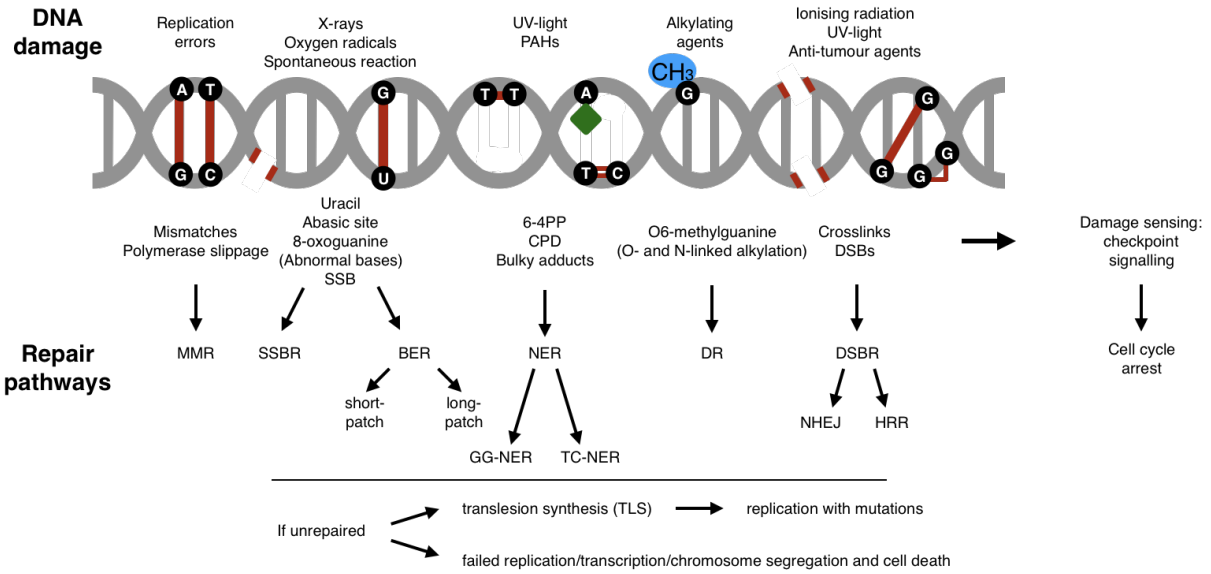


Figure 1.1: A schematic depiction of different types of DNA damage with a list of sources that cause it, as well as corresponding DNA repair pathways. Adapted from Tasaki et al. 2018.

1.1.4 DNA repair pathways

Repairing various types of DNA alterations reviewed in the previous section requires a broad range of DNA repair pathways. In total, over 80 enzymes are directly involved in DNA repair, along with over two hundred proteins being indirectly involved via regulation and recruitment of DNA repair components (Alberts et al. 2007).

DNA repair in most instances starts with damage sensing via DNA damage response enzymes or the dedicated components of the respective pathways. Upon detecting the damage, different DNA repair mechanisms are deployed depending on the scale of damage, e.g. whether it involves a single DNA strand, or is a double-strand break. If the damage is excessive or persists for too long, DNA damage checkpoints may trigger senescence or apoptosis as a last resort to preserve the integrity of the organism.

DNA damage response

After recognition of DNA damage by the enzymes capable to bind to DNA ends or adducts, *DNA damage response* (DDR) is triggered by activating the phosphatidylinositol 3-kinase-like protein kinases (PIKKs) ATM, ATR and DNA-PKs (other DNA-dependent protein kinases), or the poly(ADP)ribose polymerase (PARP) family proteins (Ciccia and Elledge 2010).

ATM and DNA-PKs are sensitive to the agents causing double-strand breaks. ATM subsequently activates the Chk2 kinase to transfer the signal downstream in the DDR

pathway. ATR is mostly recruited to the single-stranded DNA regions at stalled replication forks and DSBs, and has Chk1 serving as its signal transducer. Chk1/2 kinases, in turn, activate DNA repair factors and TP53, which regulates the expression of multiple factors affecting growth arrest, cell cycle arrest and apoptosis (Jackson and Bartek 2009). PARP proteins can be activated by both single- and double-stranded breaks, and their function is to recruit DDR factors to chromatin at DNA breaks (Ciccia and Elledge 2010).

The primary function of damage signalling proteins is to trigger the signalling cascade leading to cell cycle arrest, to allow more time for damage repair or induce apoptosis when the damage is excessive (Giglia-Mari, Zotter, and Vermeulen 2011). They mostly phosphorylate and thus activate other DNA repair enzymes or effectors – the proteins that can regulate the expression of the proteins triggering senescence or apoptosis (Giglia-Mari, Zotter, and Vermeulen 2011). DDR genes, especially effectors such as *TP53* and *PTEN*, are the most often mutated genes across all cancer types (Knijnenburg et al. 2018). Germ line defects in these genes predispose individuals to all types of cancer (Jackson and Bartek 2009).

Single strand repair

Lesions that only involve one strand of DNA can be repaired by four DNA repair pathways: direct repair (DR), base excision repair (BER), nucleotide excision repair (NER) and mismatch repair (MMR).

Direct repair. Most of the damage types require excision of the damaged base or nucleotide, but several types of damage can be repaired by *direct reversal*, or direct repair. Single-stranded nicks that only involve a break of phosphodiester bond without damage to 5' phosphate or 3' hydroxyl groups can be ligated back by DNA ligases. The Ada enzymes in *E. coli* can remove alkyl groups from O4-alkylthymines and O6-alkylguanines. Many organisms, from bacteria to animals, possess an ability to reverse cyclobutyl dimers induced by UV-light exposure via photoreactivation – a light-dependent enzymatic reaction resolving the dimer back to its original state.

Humans, however, lack photoreactivation (Li, Kim, and Sancar 1993), and only have two major types of proteins conducting direct repair: O6-methylguanine-DNA methyltransferase (MGMT) which repairs alkylation of O6-methylguanines, and ALKBH family Fe(II)/α-ketoglutarate dioxygenases (FeKGDs) that can directly reverse adenine and cytosine methylation damage in DNA and RNA, such as 3-methylcytosine. Other damage to human DNA has to be processed via more complex repair pathways.

Base excision repair. Damaged nucleotide bases in cases when the damage is relatively minor are repaired by a mechanism called *base excision repair*. It is a well-conserved pathway that strongly relies on the activity of various DNA glycosylases with different

specificities. In most organisms, BER can repair small lesions resulting from oxidation and deamination of different bases, such as 5-hydroxycytosine, 8-oxoguanine, thymine glycol, uracil (deaminated cytosine) and hypoxanthine (deaminated adenine), as well as a number of methylated bases such as 3-methyladenine, 7-methylguanine, and 2-methylcytosine (Robertson et al. 2009, Wilson III and Bohr 2007). In humans, a total of 11 different DNA glycosylases capable of processing damaged nuclear DNA have been identified (Krokan and Bjørås 2013).

Typically, base excision repair involves removal of one or several damaged nucleotides creating an abasic site, which is then converted into a single-stranded break by the apurinic/apyrimidinic endonuclease (APEX1), followed by gap-filling via DNA re-synthesis.

A sub-part of BER that deals with ligating single-strand breaks (SSBs) is often referred to as a separate SSB repair system (Caldecott 2008).

BER functions in two modes depending on the scale of the damage. The dominant pathway is normally the short-patch BER, which only affects a single nucleotide. Long-patch BER results in the creation and filling of a 2-10 bp long gap and is thought to be the dominant post-replicative BER pathway in dividing cells (Robertson et al. 2009, Krokan and Bjørås 2013).

Defects in base excision repair are often involved in disease and carcinogenesis. Biallelic loss or silencing of *MBD4*, which encodes a DNA glycosylase protecting from 5-methylcytosine deamination damage, leads to extremely high rates of C>T transitions at CpG sites and predisposes to leukemias (Sanders et al. 2018), and is often occurring in MMR-deficient colorectal cancers (Tricarico et al. 2015a). Deficiency in OGG1, NTHL1, NEIL1, MUTYH glycosylases recognising oxidative damage leads to an excess of C>A mutations (due to replication over lesions such as 8-oxoguanine) and a higher risk of colon cancer (Krokan and Bjørås 2013).

Nucleotide excision repair. A more complex and versatile DNA repair pathway is *nucleotide excision repair* (NER). Being a truly multipurpose DNA repair mechanism, it can repair a range of lesions including photodimers, DNA structure distorting lesions and bulky adducts introduced by mutagens (Nouspikel 2009). Over 25 polypeptides are involved in NER, most of which are evolutionary conserved across different organisms.

NER possesses two subsystems which can process the damage in untranscribed and transcribed regions of the genome, respectively. The first system is called *global genome* NER (GG-NER). Its main component is the XPC-HR23B protein complex, which is constantly scanning the genome of a eukaryotic cell in search of structural DNA modifications (Nouspikel 2009). Upon encountering DNA damage, TFIIH is recruited, which opens a denaturation bubble and employs XPF and XPG endonucleases to either side of the lesion to excise a 24-32 base-pair long stretch of nucleotides. The gap is then filled by PCNA

in combination with replicative DNA polymerase δ, and chromosomal nicks are sealed by XRCC1 and DNA ligases (Petruseva, Evdokimov, and Lavrik 2014).

Another modality, *transcription-coupled* NER (TC-NER), is recruited to a stalled DNA-dependent RNA polymerase (typically RNA polymerase II). Specialised TC-NER enzymes, including CSA and CSB, perform the translocation of the RNA polymerase, and the further repair proceeds the same way as GG-NER (Hanawalt and Spivak 2008).

TC-NER only repairs the damage on the transcribed (non-coding) DNA strand; hence, when both GG-NER and TC-NER are active, a transcriptional strand bias may be observed, meaning that the damage on the transcribed strand is removed more efficiently than the damage occurring on the untranscribed strand. This imbalance was observed in cells treated with UV and tobacco smoke metabolites, as well as in human cancers associated with these exposures (McGregor et al. 1991, Denissenko et al. 1998, Hollstein et al. 1991).

Germline defects in NER components have been long known to lead to very severe disease phenotypes such as xeroderma pigmentosum, Cockayne syndrome and trichothiodystrophy (Boer and Hoeijmakers 2000).

Mismatch repair. Fidelity of replication is ensured via multiple mechanisms. The replicative polymerases Pol ε and Pol δ possess high selectivity against mismatches as well as a 3'-exonuclease activity, which allows the polymerases to remove the last inserted nucleotide if it caused a mismatch. Nevertheless, they still have an error rate of about 10^{-5} to 10^{-4} depending on the nucleotide, local chromosomal properties and DNA polymerase (Kunkel and Bebenek 2000). Hence, they are backed up by an additional back-up repair pathway – *mismatch repair*.

The recognition of mismatches is typically executed by MutS protein complexes, first found in bacteria (Hsieh and Yamane 2008). In eukaryotes, there are two complexes, MutS-α and MutS-β, which have different substrate specificity due to structure of their mismatch binding sites: MutS-α preferentially detects base-base mismatches and short 1-2bp indels, and MutS-β handles larger indels up to 15 bp (Drummond et al. 1995, Habraken et al. 1996, Genschel et al. 1998). MutS proteins form a clamp sliding along the genome, which is activated upon an encounter with a mismatch, and loads the other repair complex MutL to license excision.

The MutS/MutL complex then slides away from the mismatch in order to generate single-stranded nicks on the nascent DNA that will initiate the DNA repair (Gradia et al. 1999, Kadyrov et al. 2006). The DNA strand containing a mismatch is normally removed by EXO-1 exonuclease, and the resulting gap is filled by lagging strand polymerase δ (Goellner, Putnam, and Kolodner 2015).

The MMR system should be exclusively removing the mismatches from the newly

synthesised strand. The strand recognition mechanism is not clear for most organisms, but the leading candidate is PCNA: it has to be recruited to the repair site to enable the endonuclease activation, but it can also discriminate between the new and the old strands based on its loading orientation (Pluciennik et al. 2010).

Germline defects in MMR genes cause conditions associated with a high risk of cancer development, and somatic alterations are also often found in tumours (Bonneville et al. 2017).

Translesion synthesis. If DNA damage is encountered during DNA synthesis, it has to be overcome by the mechanism called *translesion synthesis* (TLS). A specialised set of DNA polymerases is capable to adapt modified DNA as a template and perform the synthesis by inserting a base opposite the lesion: Y-family polymerases η, κ, ι (Sale, Lehmann, and Woodgate 2012), B-family polymerases ζ and REV1 (Gan et al. 2008), and A-family polymerase θ (Seki, Marini, and Wood 2003).

TLS polymerases have more flexibility in templates, but are less accurate and have no proofreading ability, therefore their error rates reach 10^{-4} to 10^{-1} (Sale, Lehmann, and Woodgate 2012). Hence, TLS polymerases provide the ability to tolerate the damage and avoid replication failure (which can lead to DSBs and cell death) at the cost of introducing more mutations. Translesion synthesis typically occurs during replication but can also occur as a part of other repair mechanisms, such as cross-link repair when a gap-filling synthesis across damaged DNA is required (Sale 2013).

Many of the TLS polymerases are interchangeable, however, they have different efficiency and error rates when replicating over different lesions. Defects in polymerase η (also termed XPV) leads to a xeroderma pigmentosum phenotype without the defects in NER machinery itself, indicating that it is essential to tolerating UV damage, and cannot be replaced by other polymerases (Masutani et al. 1999).

Of other TLS polymerases, polymerase κ has a preference for correct synthesis over G adducts and extends the synthesis from other lesions. REV1 is controlling the TLS induction, polymerase ζ and ι can deal with a range of damaged bases but with an increased error-rate. X-family TLS polymerase β, μ and λ take part in BER and NHEJ, and TLS polymerase θ is an important component of crosslink repair and double-strand break repair (Sale 2013). In mammals, TLS polymerases η, ι and REV1 play a crucial role in somatic hypermutation (Zeng et al. 2001, Faili et al. 2002, Masuda et al. 2009).

Double-strand break repair

Double-strand breaks are typically the most lethal type of DNA damage and may result in cell death. Consequently, there is more than one repair pathway that serves to repair DSBs. Depending on the cell cycle phase and cell conditions (e.g. damage intensity

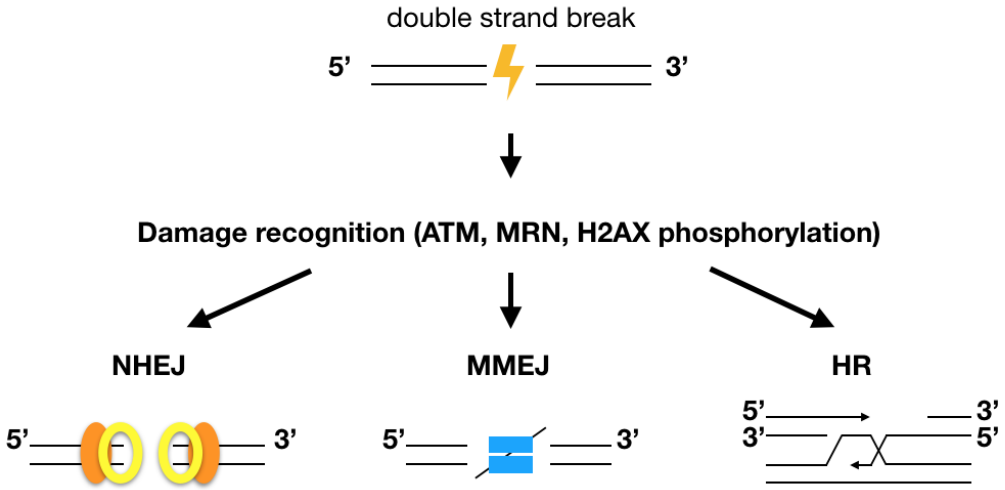


Figure 1.2: A scheme of pathway choices for double-strand break repair. Adapted from Kim, Hromas, and Lee 2013.

and repair enzyme availability), there are three mechanisms that can complete this task (Figure 1.2).

Homologous recombination repair. The preferable pathway for DSB repair is *homologous recombination repair* (HRR), which uses the sister chromatid as a template to restore the information that may be lost around the double-strand break (Boulton 2010). Typically, it is triggered when a DSB is detected in late S or G2-phase (Jackson 2002). HRR is capable of accurately repairing the majority of DSBs. Moreover, recombination repair is an important mechanism performing the crossing-over during meiosis (Jackson 2002).

There are currently four models of HRR: classical double-strand break repair (cDSBR), synthesis-dependent strand annealing (SDSA), break-induced replication (BIR) and single-strand annealing (SSA) (Li and Heyer 2008). All of them share the first essential steps (Figure 1.3).

HRR starts with the MRN complex binding to the DNA on either side of the DSB, tethering the ends and cleaving nucleotide links as well as signalling to recruit other repair components. After that, 5' ends are resected, and the double-stranded DNA is opened by the helicases. The resulting single-stranded DNA is then cut by the exonucleases and coated with RAD51 protein to protect it from endogenous damage. Then, a search for a homologous DNA template is conducted by aligning the sister chromatid. Upon finding one, the strand invasion begins and a displacement-loop is created, which results in the formation of a heteroduplex. Then, a DNA polymerase is recruited to synthesise the missing DNA, and the D-loop changes into a cross-shaped structure termed Holliday Junction (Figure 1.3).

In classical DSBR, the invasion will happen for both strands at the same time (two-end invasion), forming two Holliday junctions. This mechanism is also required for meiosis to ensure the formation of crossover products (Figure 1.3). In another repair option, SDSA, only one strand is synthesised using the homologous template, after which it is displaced and annealed, while the gap on the other strand is filled by complementary DNA synthesis. Thus, SDSA leads to non-crossover products only, but happens in both mitotic and meiotic cells (McMahill, Sham, and Bishop 2007).

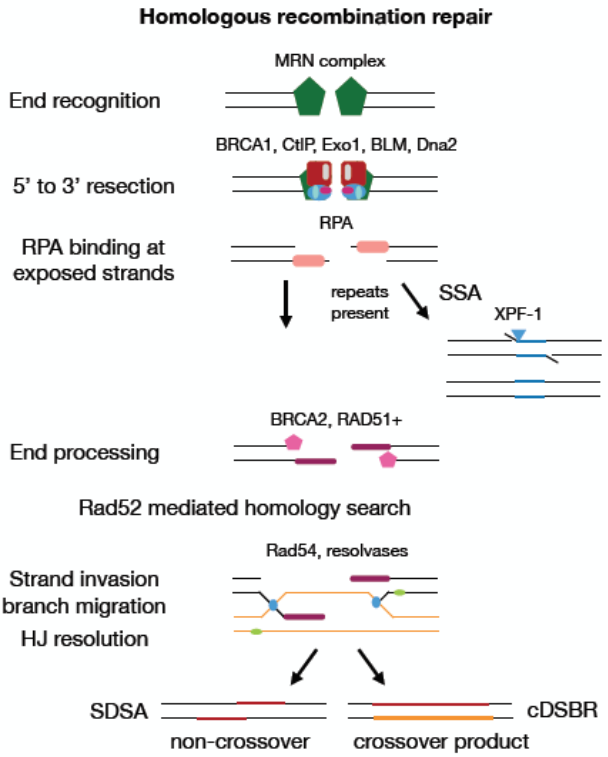


Figure 1.3: A scheme of homologous recombination repair steps and model choice.

Break-induced replication is usually recruited to repair DSBs created during the fork collapse upon replication stress (Kramara, Osia, and Malkova 2018). DSBs in one of the two newly created dsDNA regions of the replication fork are essentially one-sided (only the break in the old strand needs to be repaired), hence a single-end invasion to the homologue is required: the broken end invades the homologous sequence, and unidirectional DNA synthesis is initiated from the invasion site. This structure can replicate several hundred kbps after which it separates. If needed, re-invasion and synthesis are triggered again until the repair is complete. This type of repair can create genomic rearrangements and is typically avoided by the cells unless it is necessary to restart replication (Deem et al. 2011).

If the break happened in a repetitive sequence, the resected ends can be repaired by SSA (Figure 1.3). In this case no invasion is needed: direct repeats uncovered during resection will be annealed together and XPF/ERCC1 will cleave the flapping strand ends. As it only relies on connecting the repeats, SSA is the most mutagenic type of homologous recombination repair (Bhargava, Onyango, and Stark 2016).

Non-homologous end-joining. If the HRR is not available or too slow, another DSB repair pathway takes over: *non-homologous end-joining* (NHEJ). NHEJ is a cheap yet mutagenic way of repairing DSBs as it simply processes and ligates the ends of a DNA double helix (Figure 1.4). In diploid organisms, it typically occurs during early S or G0/1 phases when the homology donor is not nearby (Lieber 2010). In somatic non-dividing

cells, NHEJ is the dominant DSBR repair pathway (Chakraborty et al. 2016).

NHEJ requires the Ku70/80 complex, which forms a subunit of DNA-dependent protein kinase (DNA-PK) (Figure 1.4). This complex binds to DNA ends, bridging them and protecting from non-specific nucleases, and triggers the assembly of NHEJ complex involving DNA-PKs and a set of end-processing enzymes, which then perform terminal end processing. Finally, once the blunt ends are created, XRCC4/DNA ligase 4 complex ligates DNA ends, and then the NHEJ complex dissolves (Davis and Chen 2013).

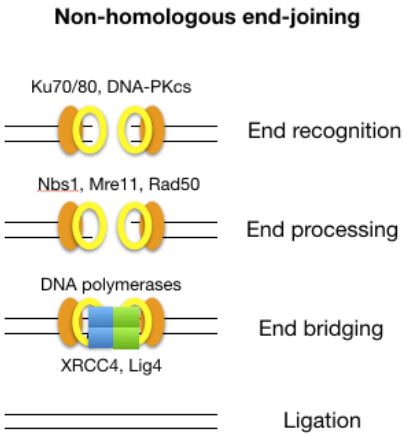


Figure 1.4: A scheme of non-homologous end-joining repair.

If the overhanging DNA ends match exactly, this repair process will not introduce mutations (Rodgers and McVey 2016). NHEJ also shows high accuracy when repairing the breaks detected by RNA polymerase II in coding regions (Chakraborty et al. 2016). Otherwise, it can be mutagenic by introducing translocations (if the ends from different breaks get ligated together) or small insertions/deletions during the processing and ligation of DNA break ends.

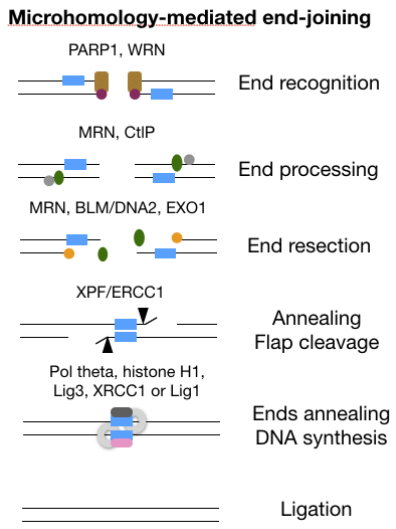


Figure 1.5: A scheme of microhomology-mediated end-joining repair.

Alternative end-joining. In the absence of the core components of NHEJ, or if their recruitment is delayed, a slower yet efficient alternative end-joining mechanism can operate in DSB repair (Figure 1.2) (Chang et al. 2017). *Alternative end-joining* (a-EJ) pathway, also known as *microhomology-mediated end-joining* (MMEJ) or *Pol θ mediated end-joining*, was first identified as a mechanism relying on the 2-20 bp long homologous sequences at either side of the break (Figure 1.5). Based on the level of homology required for the repair, MMEJ was put in between NHEJ, which processes the ends of a break until they have 1-4 complementary bases, and SSA which typically needs over 20 bp of homology (hence it works well in repetitive regions) (Bhargava, Onyango, and Stark 2016). HRR, in comparison, requires over 100 bp of homologous sequence to consider it a template for repair (Lisby and Rothstein 2015).

In fact, a-EJ does not necessarily require microhomology, but MMEJ happens with much higher frequency than microhomology-independent repair (Mansour, Rhein, and

Dahm-Daphi 2010). Both, however, require the same basic steps to preform DSB repair (Figure 1.5). DSB detection starts with the PARP1 protein binding to the exposed DNA ends. Then the MRN complex along with the CtIP resection factor, BLM helicase, and EXO1 exonuclease, perform end resection. The homologous regions on either side of the break are annealed together (or single bases, if there is no microhomology), and flapping ends are resected by XPF/ERCC1 factor. Finally, the gaps are filled by polymerase θ, an essential component of the pathway that can extend mismatched termini, and the DNA is ligated by the ligation complex. MMEJ typically occurs in S-phase only (Seol, Shim, and Lee 2018).

MMEJ plays a major role in DSB-induced mutagenesis: one of the homologous regions and the sequence between them are always deleted (Schendel et al. 2016). MMEJ has also been suggested to take part in the formation of chromosomal translocations and large-scale rearrangements, as the breakpoints from chromosomal translocations in somatic cells often have some degree of microhomology (Schimmel et al. 2017).

Crosslink repair

DNA crosslinking occurs when two non-pairing DNA bases form a covalent bond between them. Typically, it happens via a reaction with a molecule capable of covalently binding to two DNA bases at the same time (Noll, Mason, and Miller 2006).

DNA crosslinks may involve nucleotide on the same strand (intrastrand crosslinks), in which case they are repaired by NER (O'Donovan et al. 1994), or bases on the opposite strands (interstrand crosslinks, ICL). Interstrand crosslinks have a more serious effect on DNA structure, and can easily lead to double-strand breaks. Two mechanisms of ICL repair have been observed in eukaryotes: recombination-dependent and recombination-independent (Huang and Li 2013).

Recombination-independent, or mutagenic, interstrand crosslink repair mainly occurs in G1 phase or in quiescent cells. It requires enzymes from multiple pathways. The crosslink is recognised by NER enzyme XPC, and there is evidence that ICLs highly affecting the DNA structure can also be detected by MMR enzymes (Kato et al. 2017). First, a gap is created in one of the strands by creating nicks on either side of the crosslink (a process termed "unhooking"), leaving the crosslinked base attached to the intact strand. The gap is then filled by a translesion synthesis polymerase, which can insert incorrect bases. Then, the monoadduct on the other strand is removed by NER with a strand incision and gap filling on the other strand. This can lead to mutations due to errors made by a TLS polymerase on the first strand (Zheng et al. 2003). When abasic sites occur as intermediates during ICL repair, BER enzymes cover them to protect from reacting with other bases or proteins, and to avoid the formation of additional crosslinks.

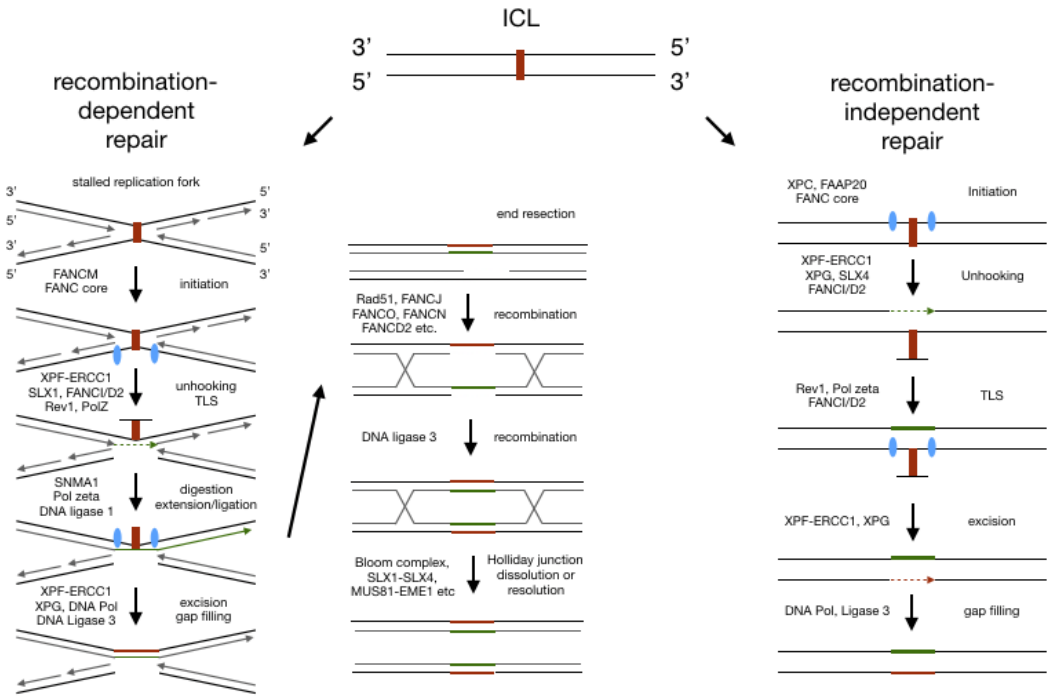


Figure 1.6: A scheme of the mechanisms repairing interstrand crosslinks. Adapted from Huang and Li 2013.

Recombination ICL repair, also termed replication-dependent repair, happens in late S or G2-phase when ICLs are detected by the replication fork stalling. According to the current model (Hashimoto, Anai, and Hanada 2016), ICL causes stalling of two replication forks on both sides of it. Similar to the first mechanism, NER structure-specific nucleases unhook the ICL by making dual incisions on either side of the ICL on one strand, the resulting gap filled by TLS polymerases that bypass the remaining lesion and connect to the Okazaki fragments. The crosslinked oligonucleotides are then repaired by NER factors. Finally, a DSB on the sister chromatid is repaired via a classical HRR pathway using the restored duplex DNA as a template.

This complex mechanism is regulated by a set of proteins which were identified by studying patients with a particular form of genetic bone marrow disease, the Fanconi anaemia (FA) proteins (Lobitz and Velleuer 2006). The FA-mediated signalling is crucial for the recruitment of incision factors and transition into HRR (Huang and Li 2013).

The disease that gave the name to the FA pathway was associated with bi-allelic germline defects in these genes, which also included several HRR related genes such as *BRCA1/2*, *PALB2*, *RAD51C*, *XRCC2* and TLS component *REV7* (Dong et al. 2015) which are essential for successful crosslink repair. Individuals affected by this disease very often develop bone marrow failure, developmental abnormalities, and cancer (Lobitz and Velleuer 2006).

1.2 Mutational signatures

In the previous section, I introduced the types of DNA damage and repair, which commonly occur in the cells and generate mutations. The genome of a cell contains the mutational footprints of different genotoxic exposures experienced by the lineage. Various mutational processes and repair mechanisms discussed in the sections above lead to characteristic distributions of mutations. These are called *mutational signatures*. The deconvolution of the mutational traces, or signatures, of individual processes can shed light on potential sources of pathogenic mutations that led to cancerous transformation. In this section, I will review the recent findings in the field of mutational signature analysis and describe their clinical significance for cancer research.

1.2.1 Somatic mutations in cancer

Cancer is a set of diseases associated with the uncontrolled growth of cells. In the majority of cases, this increased proliferation rate stems from genetic alterations in cancer genes, which activate oncogenes or knockout tumour suppressor genes (Stratton, Campbell, and Futreal 2009). According to Armitage and Doll's multistage theory of carcinogenesis, there are typically around 6 mutational events that lead to the disease onset (Armitage and Doll 1954).

Recent studies based on the whole-genome sequencing of a large number of tumours demonstrated the prevalence of clonal expansions and ongoing evolutionary processes in cancer (Greaves and Maley 2012). Somatic mutations aggregated in tumour cells can shed light on the genomic history and evolution of the tumour (Nik-Zainal et al. 2012b, Gerstung et al. 2018). This evolution may be gradual or punctuated by catastrophic events such as chromothripsis (Forment, Kaidi, and Jackson 2012). Over the course of cancer evolution, the mutation rate of all cells or different clones can change due to the temporal activity of different mutational processes, or somatic disabilities in DNA repair pathways occurring at different time points (Greaves and Maley 2012).

Mutations found in cancer cells may be either driver mutations if they alter the fitness of the cell and provide it a selective advantage, or passengers if they do not affect the fitness but happen to be carried by the same cells which have acquired a beneficial mutation (Stratton, Campbell, and Futreal 2009). The set of driver mutations is typically small (Martincorena et al. 2017, Tomasetti et al. 2015). The majority of them tend to affect the genes essential for cell growth or damage-caused apoptosis such as tumour suppressor gene *p53* (Rivlin et al. 2011) or oncogene *KRAS* (Wang et al. 2015).

The collection of driver and passenger mutations is a consequence of the mutational processes which were active in a given tumour. Hence, the overall distribution of mutations

acquired at different times contains information about the mutagenic processes which happened in the past. Unique spectra of mutations generated by different mutational processes were termed *mutational signatures*. Analysing the mutational signatures of the individual and interaction effects of key elements of genome stability maintenance and environmental exposures and finding them in the mutational spectra of a tumour provides the means to trace back the causes of the disease and predict the outcomes of chemo- or radiotherapy for individual cases (Poon et al. 2014).

Mutational signatures are not only applicable to cancer studies. Aggregation of germ line data has allowed the examination of the distribution of mutations contributing to human variation and genetic diseases (Chen, Férec, and Cooper 2013). Additionally, the introduction of high-depth sequencing opened the way to studying the processes creating somatic heterogeneity in healthy tissues (Behjati et al. 2014). Recent studies have demonstrated a frequent and age-dependent presence of clones with driver mutations and ongoing mutational processes in healthy tissues (Martincorena et al. 2015, Yadav, DeGregori, and De 2016, Martincorena et al. 2018).

1.2.2 Signatures of mutational processes active in human cancer

Somatic mutations in cancer genomes are generated by multiple endogenous and exogenous processes with unknown timing and intensity. The genome of a tumour at the time of diagnosis presents a superposition of these processes. Being able to deconvolve the mutational spectra of different tumours into the mutational signatures would allow inferring the processes which shaped these spectra (Figure 1.7). The recent explosion in the amount of publicly available cancer sequencing data has made it possible to computationally identify and characterise the patterns of mutations present in human cancers (Alexandrov et al. 2013b).

The decomposition of the mutational spectra into contributing factors currently has several assumptions. Firstly, to be detected with confidence, each process should affect at least several samples. Secondly, the mutational spectra corresponding to these factors are assumed to be constant across samples. Finally, different processes are thought to act independently and contribute mutations in an additive manner. Under these assumptions, one would expect that each of the mutational processes shaping cancer genomes would correspond to a unique pattern of mutations, which can be identified from factor analysis of many tumours.

The initial list of mutational signatures in cancer considered only single-base substitutions (SBS) (Alexandrov et al. 2013b). The most recent version also inferred signatures from dinucleotide variants (DNVs) and small indels (Alexandrov et al. 2018). According

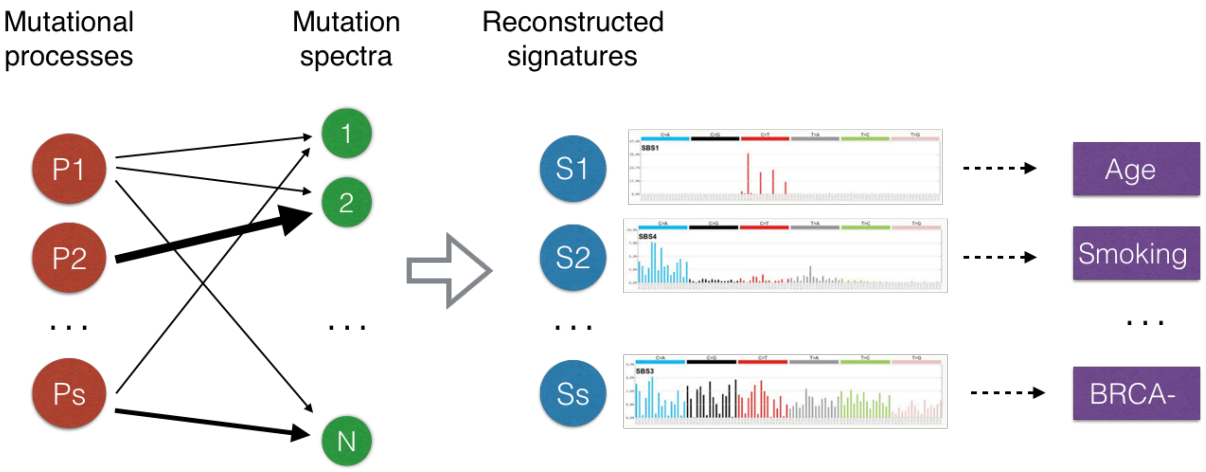


Figure 1.7: The concept of mutational signature analysis. Underlying mutational processes P_1, .., P_S shape the mutational spectra of N samples, from which one extracts S computational signatures and associates them to the mutagenic factors.

to the current state of the Catalogue Of Somatic Mutations In Cancer (COSMIC) project (Forbes et al. 2015, Alexandrov et al. 2018), there are over 60 single-base substitution, 11 dinucleotide and 17 indel signatures as identified based on several thousand cancer genomes and exomes (Weinstein et al. 2013, Campbell et al. 2017). These signatures are largely defined by the relative frequency of different mutation types based on pyrimidine reference and the local sequence context (adjacent 5' and 3' base for SBS, length of the repeat for indels in repetitive regions).

It is important to note that these signatures may not be unique; they were extracted from a large dataset of cancer genomes as one possible basis of the latent variable space, and their linear combination may represent more meaningful signatures (e.g., as was proven for signatures 1A, 1B and 5 in Alexandrov et al. 2015), especially as soon as more data is added to the pool, increasing statistical power. In addition, signature identification is further complicated by high interpersonal variability and the potential presence of mutational signatures with similar mutation spectra (Baez-Ortega and Gori 2017). Most importantly, the mutational signatures identified computationally do not necessarily have a biologically relevant cause.

1.2.3 Methods for learning mutational signatures from mutational spectra

The investigation of latent factors which induce mutations in cancer genomes has gained much attention recently. Several methods have been developed for the *de novo* extraction of mutational signatures, as well as for the decomposition of mutation spectra

over a set of known signatures (or signature re-fitting).

Formulated in mathematical terms, the signature extraction task turns into a blind source separation problem (Comon and Jutten 2010). Every sample i (out of N samples) is characterised by a mutation spectrum y_i, which is a set of counts for M different mutation types, $y_i = \left(y_i^1, ..., y_i^M\right)$, $y_i^k \in (\mathbb{N} \cup \{0\})$. Assume there are P mutational processes s_j defined by the probability distribution of acquiring mutation of a particular class $s_j = \left(s_j^1, ..., s_j^M\right)$, $s_j^k \geq 0$, $\sum_{k=1}^{M} s_j^k = 1$. Then every spectrum can be approximated (down to the noise) as a linear combination of processes:

$$y_i = \sum_{j=1}^{P} e_i^j s_j + \epsilon_i, e_i^j \geq 0$$

where $e_i^j, j = 1, ..., P$ are the exposures which quantify the contributions of mutational processes to the sample i, and ϵ_i is the random noise added to the spectrum.

In matrix form, it can be represented as a matrix factorisation problem:

$$Y = E \times S^T + \epsilon, \tag{1.1}$$

where Y is a $N \times M$ matrix of mutation counts per sample, E – a $N \times P$ exposure matrix, S – a $M \times P$ matrix of mutational signatures, and ϵ – a random error matrix. Hence, signature extraction methods aim to find such non-negative exposure and signature matrices that their product is close enough to the original count matrix under some feasible assumptions for the noise and biological interpretation of the factors. There are several machine learning and statistical techniques that can be applied to solve this task.

Non-negative matrix factorisation

The most common technique for mutational signature analysis is non-negative matrix factorisation (NMF), which was initially introduced for extracting latent factors in image analysis (Lee and Seung 1999). In contrast to other dimensionality reduction techniques, such as principal or independent component analyses, NMF enforces non-negativity of the components and their mixture coefficients yet makes no assumptions about orthogonality or independence of the latent features. NMF has been successfully applied to mine large biological data for patterns such as gene expression signatures (Brunet et al. 2004, Devarajan 2008).

The initial NMF algorithm suggested for signature extraction in Alexandrov et al. 2013a aimed to minimise the Frobenius norm of the difference between the real and reconstructed count matrix. This was equivalent to assuming a normal distribution for

the noise:

$$Y \sim \mathrm{N}(E \times S^T, \sigma^2),$$

$$(E, S) = \mathrm{argmin}_{E,S} \left(||Y - E \times S^T||_F \right)^2,$$

$$||Y - E \times S^T||_F = \sqrt{\sum_{i,j} (y_{i,j} - \sum_{k=1}^{P} e_i^k s_k^j)^2}. \tag{1.2}$$

An important drawback of NMF is the necessity to manually choose the most optimal number of signatures, P, which is often chosen based on the stability of signatures (or cophenetic distances between them) and the quality of reconstruction (Alexandrov et al. 2013a).

As Gaussian noise does not best describe the underlying variation in the data, several successive methods (Alexandrov et al. 2018) switched to minimising Kullback-Leibler divergence as an objective, which was equivalent to assuming a Poisson generative model for the observed data:

$$Y \sim \mathrm{Poisson}(E \times S^T),$$

$$(E, S) = \mathrm{argmin}_{E,S} \left(D(Y||E \times S^T) \right),$$

$$D(Y||E \times S^T) = \sum_{i,j} \left(y_{i,j} \log \left(\frac{y_{i,j}}{\sum_{k=1}^{R} e_i^k s_k^j} \right) - y_{i,j} + \sum_{k=1}^{R} e_i^k s_k^j \right). \tag{1.3}$$

Recently, a penalised version of NMF was suggested to enforce the sparsity of the factors (Ramazzotti et al. 2018) but the biological relevance of this assumption still has to be determined.

Following the Bayesian NMF approach proposed by Cemgil 2009, another probabilistic NMF model was developed by Fischer et al. 2013. They calculate E and S matrices using an expectation-maximisation algorithm which explicitly assumes a Poisson distribution for the mutation counts. The latent factors are assigned a Gamma prior, which essentially corresponds to assuming negative-binomial distribution for the observed counts.

Other probabilistic interpretations of NMF were also suggested: bayesNMF (Kasar et al. 2015) and SignatureAnalyzer (Alexandrov et al. 2018) use a Bayesian framework to specify the priors on the matrix factors, and also apply the automatic relevance determination to automatically choose the best number of signatures. signeR (Rosales et al. 2016) uses an empirical Bayes approach to NMF, takes into account the tumour-specific opportunity matrix and employs Markov Chain Monte-Carlo sampling to obtain the samples

from posterior distributions of the factors.

Bayesian approaches and Dirichlet processes

Despite NMF being the most popular method for signature extraction, there are other approaches based on the methods initially developed for natural language processing. The most suitable of them is topic modeling, which aims to infer latent factors, or topics, of each word in a set of documents drawn from the same vocabulary (Blei and Lafferty 2007). Topic modeling for mutational signature analysis interprets observed mutational spectra as documents, mutation types as vocabulary, and the signatures as topics.

Most topic modeling methods are based on Bayesian models such as latent Dirichlet allocation (LDA) (Blei, Ng, and Jordan 2003). pmsignatures (Shiraishi et al. 2015), a method which considers per-position mutation distributions rather than per-type, uses a generalised LDA. A method based on Dirichlet processes was developed, which imposes a prior on the number of signatures and uses hierarchical Dirichlet processes to model different numbers of signatures in different tissues and samples (Li et al. 2017b). To combine signatures of different mutational classes, Funnell et al. 2019 created a correlated topic model that extracted signatures from the base substitution counts as well as copy number profiles, and used the correlation structure between them to better assign signatures to samples. This approach can also be used to specify the effect of clinical or genetic variance on the exposures (Robinson, Sharan, and Leiserson 2019).

A recent comparison of performance across different *de novo* signature analysis methods suggested that probabilistic methods are superior over NMF-based ones in correctly reconstructing signatures used to generate the simulated dataset (Omichessan, Severi, and Perduca 2019).

Refitting approaches

Apart from searching for signatures, one may want to decompose a set of samples over a known set of signatures. This task is usually referred to as refitting, as opposed to *de novo* signature extraction. Some studies also point out that the exposure estimates are more stable if they are obtained via a separate refitting procedure and not simultaneously with signature extraction (Alexandrov et al. 2018).

Rosenthal et al. 2016 presented an iterative multiple linear regression approach to fit signatures S to the sample matrix Y. This work was extended by Huang, Wojtowicz, and Przytycka 2017 and Krüger and Piro 2019 to ensure the optimality by the use of quadratic programming and simulated annealing. In addition, non-negative least squares can be used to solve this task (Blokzijl et al. 2018), as well as Bayesian inference (Gori

and Baez-Ortega 2018) and hidden Markov models (Wojtowicz et al. 2019).

1.2.4 Associating signatures with their origins

Typically, the first and easiest approach to trace the causal factor behind a mutational signature is visual investigation. Mutational properties of a number of genotoxic agents have been well established, and a preference for a particular type of mutation in the signature may be indicative of the underlying mutagenic process, such as in CpG>TpG changes for spontaneous deamination of 5-meC or T>A mutations for aristolochic acid exposure. Alternatively, a wider distribution of mutations may suggest a link to a more complex genotoxin or an inactivation of a DNA repair pathway. However, mutational spectrum can only suggest a set of possible candidates, and one still needs to establish association between a factor and a signature, and then prove the causative role of this factor.

As signature extraction is usually performed in an unsupervised manner, there is no guarantee that the extracted signatures have a biological meaning. Based on the cancer types where a signature occurs, particular factors are expected to be involved in the mutagenesis, such as UV-light exposure in skin cancers or MMR deficiency in colorectal cancers (Armstrong, Kricker, and English 1997, Bronner et al. 1994). Currently, the most common approach to linking a signature with its underlying cause is to find a statistically significant association between the exposure to this signature and a clinical, molecular or demographic feature.

Thus, several correlation analyses were conducted that suggested associated mutational mechanisms for about a half of the single-base substitution signatures found in cancers (Table 1.2) (Helleday, Eshtad, and Nik-Zainal 2014, Alexandrov et al. 2018). Two omnipresent signatures were associated with the age of the patient (Alexandrov et al. 2015), and exposures to many carcinogens such as UV-light, smoking, aristolochic acid, aflatoxin, haloalkanes, and chemotherapy drugs were correlated each with a respective signature (Table 1.2) (Poon et al. 2014). Recently, Zhivagui et al. 2019 identified a signature present in many cancers types and potentially associated with exposure to acrylamide, which was previously shown to cause cancer in mice.

Deficiencies of different DNA repair components were found to be correlated with individual signatures, sometimes more than one: 7 signatures were associated with MMR deficiency, two of them being correlated with concurrent MMR and polymerase proofreading domain defects (Haradhvala et al. 2018). One of the most wide-spread mutational processes characterised by two signatures of C>T and C>G mutations was found to be associated with the activity of cytidine deaminases from the activation-induced cytidine

deaminase (AID)/apolipoprotein B editing complex (APOBEC) family (Nik-Zainal et al. 2012a).

A strong correlation was observed between one of the base substitution signatures frequently occurring in breast and pancreatic cancers and the defects in BRCA1/2 genes, and generally with homologous recombination deficiency (Alexandrov et al. 2013b, Polak et al. 2017, Riaz et al. 2017). Further analysis of the signatures composed of copy number changes and structural variants revealed several signatures associated with HR deficiency which correlated with different components being defective (Davies et al. 2017, Macintyre et al. 2018, Funnell et al. 2019). This analysis suggested differences in mutagenesis between BRCA1 or BRCA2-associated HRD, and other types of HR deficiency.

Despite the methodological progress, many signatures are still lacking any indication of their aetiology. About 30% of these signatures occurred in small proportions across samples from different cancer types and were proposed to be stemming from sequencing or variant calling artefacts (Alexandrov et al. 2018).

Some signatures showed associations only detectable in certain cancer types or patient groups: significant correlation was observed for the contribution of signature SBS 5 and defects in NER gene *ERCC2* in bladder tumours of smoking patients (Kim et al. 2016).

Signature	Potential cause	Validation
SBS 1	ageing Alexandrov et al. 2015	Blokzijl et al. 2016
SBS 2/13	activity of APOBEC family of enzymes (Nik-Zainal et al. 2012a)	Schumacher, Nissley, and Harris 2005
SBS 3	biallelic BRCA1/2 defects (Alexandrov et al. 2013b, Polak et al. 2017, Riaz et al. 2017)	Zou et al. 2018
SBS 4	tobacco smoking Alexandrov et al. 2016	Kucab et al. 2019
SBS 5	clock-like Alexandrov et al. 2015, may be associated with ERCC2 in tobacco-exposed bladder cancers Kim et al. 2016	-
SBS 6, 14, 15, 20, 21, 26, 44	MMR deficiency (Alexandrov et al. 2013a, Supek and Lehner 2015)	Drost et al. 2017, Zou et al. 2018
SBS 7a/b/c/d	UV-exposure (Alexandrov et al. 2013b, Hayward et al. 2017)	Kucab et al. 2019
SBS 8	supposedly GG-NER deficiency	Jager et al. 2019
SBS 9	activity of Pol η in somatic hypermutation (Puente et al. 2011)	-
SBS 10a/b	proofreading deficiency of Pol ϵ (Cancer Genome Atlas Network 2012, Cancer Genome Atlas Research Network et al. 2013)	Shinbrot et al. 2014

SBS 11	exposure to temozolomide (Alexandrov et al. 2013b)	-
SBS 14	concurrent MMR deficiency and Pol ϵ proofreading deficiency (Haradhvala et al. 2018)	-
SBS 16	supposedly associated with alcohol consumption (Li et al. 2018)	-
SBS 17a/b	possibly dTTP pool depletion mediated by 5-fluoracil or similar base analogue	Christensen et al. 2019
SBS 18	oxidative damage	Drost et al. 2017, Zou et al. 2018
SBS 20	concurrent MMR deficiency and Pol δ proofreading deficiency (Haradhvala et al. 2018)	-
SBS 22	exposure to aristolochic acid (Poon et al. 2013)	Poon et al. 2013, Kucab et al. 2019
SBS 24	exposure to aflatoxin (Poon et al. 2014)	Huang et al. 2017a, Kucab et al. 2019
SBS 29	tobacco chewing (International et al. 2013)	-
SBS 30	BER deficiency due to NTHL1 inactivation	Drost et al. 2017
SBS 31/35	prior chemotherapy with platinum agents (Alexandrov et al. 2018)	Boot et al. 2018, Kucab et al. 2019
SBS 32	prior treatment with azathioprine to induce immunosuppression (Alexandrov et al. 2018)	-
SBS 36	BER deficiency due to MUTYH defects (Viel et al. 2017)	Pilati et al. 2017
SBS 42	occupational exposure to haloalkanes (Alexandrov et al. 2018)	Mimaki et al. 2016
SBS 84/85	activity of AID (Alexandrov et al. 2018)	Pettersen et al. 2015

Table 1.2: Mutational signatures in cancer and their proposed aetiology.

1.2.5 Experimental validation

A correlation between a signature and a mutagenic factor is not sufficient to establish a causative relationship between them. Therefore, much attention turned to the experimental validation of signatures in model systems, which involved both re-analysing the existing genome-wide data on mutagenesis and conducting new experiments (Olivier et al. 2014).

Several types of human-based systems were recently employed to validate the concept of mutational signatures and confirm their aetiology. Zou et al. 2018 used immortalized human near-haploid cell line HAP1 and CRISPR-Cas9 gene editing to produce detailed profiles of several DNA repair knockouts including MMR, HRR and BER factors. The sequencing of genetically manipulated stem-cell derived human organoids demonstrated the presence of age-related signature and signatures associated with MMR and NER deficiencies (Blokzijl et al. 2016, Drost et al. 2017, Jager et al. 2019). Further on, Kucab et al. 2019 established the protocol for studying the accumulation of mutations in the clones derived from human induced pluripotent stem cells (iPSCs), which allowed to describe the mutational spectra of many environmental exposure agents such as UV-light, aristolochic acid, and platinum-based drugs in iPS cells.

Many mutational processes were studied in cancer cell lines. The signatures of AID and APOBEC were dissected using B-cell lymphoma cell lines (Pettersen et al. 2015). The propagation of a large set of cancer cell lines allowed to distinguish between the processes which were active in the past and hence are likely to be related to exogenous exposures from those which are still on-going, such as DNA repair deficiencies, defects of POLE proofreading domain, or endogenous damaging agents including reactive oxygen species and over-active APOBEC enzymes (Petljak et al. 2019).

However, even the results from human-based model systems are not fully reflective of the mutagenic activity that various agents have in the context of cancer. Many mutagens require metabolic activation to obtain the reactive metabolite capable of damaging DNA in the same way as in tumours (Guengerich 1992). Moreover, several experiments showed high contamination by background mutational process which can mask weak mutational patterns introduced by DNA repair deficiencies (Zou et al. 2018).

Apart from human cells, many other model systems were utilised to study signatures (Segovia, Tam, and Stirling 2015). Multiple studies from bacteria (Maharjan and Ferenci 2014, Matsumura et al. 2018), yeast (Degtyareva et al. 2019, Larsen et al. 2017), nematodes (Meier et al. 2014, Meier et al. 2018, Huang et al. 2017a), murine models (Connor et al. 2018, Dow et al. 2018) and other model systems such as chicken fibroblasts (Szikriszt et al. 2016) generated a huge amount of data that reveals the mechanisms as well as the distribution of damage of multiple genotoxic agents with respect to a model's genome.

Many model organisms have the advantage of simplicity, lower costs and shorter time scales compared to human-based systems. A study by Mimaki et al. 2016, which found a mutational spectrum in cholangiocarcinomas, presumably associated with haloalkane exposure, successfully used *Salmonella typhimurium* to validate this finding. Well-established model systems, such as *S. cerevisae* and *C. elegans*, are equipped with large and consistent libraries of knockouts and well-annotated genomes, which allows exploring the mutational

response to mutagenic exposures across several genetic backgrounds (Meier et al. 2014), as well as study a single process such as oxidative damage (Degtyareva et al. 2019) or replication fork stalling in great detail (Larsen et al. 2017). Many studies of genome-wide mutational distributions were conducted in mice, the most well-described model to study the induction and development of cancer, helping to dissect the mutagenic contributions of many mutagens (Nik-Zainal et al. 2016, Huang et al. 2017b), to explore the variability of mutational landscapes in liver tumours (Connor et al. 2018, Dow et al. 2018), as well as to describe the mutations and evolution of skin metastases (McCreery et al. 2015) and KRAS-driven lung cancers (Westcott et al. 2015).

1.2.6 Clinical applications of mutational signatures

Apart from providing valuable insights in mutagenesis mechanisms, the utilisation of mutational signatures in clinical research also provides a number of benefits.

Therapeutic biomarkers. Mutational signatures provide several types of clinically relevant information (Ma et al. 2018). Experimentally confirmed links between signatures and DNA repair deficiencies allow detecting these defects when the biomarkers or individual gene analysis may fail to give the correct indication. The presence of a DNA repair deficiency signature confirms the actual inability of the cell to repair the relevant damage. Measures of DNA repair deficiencies can be exploited as an indicator of a patient's response to treatments such as chemotherapy (Hegi et al. 2005, Helleday et al. 2008), immunotherapy (Le et al. 2017), or synthetic lethality approaches (Shaheen et al. 2011).

Cancer prevention. Many exogenous mutagenic exposures, including tobacco smoking and UV-exposure, were shown to induce mutations and lead to cancer (Pfeifer 2010, Armstrong, Kricker, and English 1997, Poon et al. 2014). Treatment with chemotherapy drugs also generates specific traces of mutations (Szikriszt et al. 2016). Analysis of signatures associated with environmental or lifestyle factors can be predictive of the therapy outcome (Trucco et al. 2019) and indicative of the aetiology of cancer (Poon et al. 2014), which can help to improve both the efficacy of treatment and understanding of the epidemiology of cancer.

Outcome prediction. Moreover, even if the precise causes of signatures are not known, stratification of patients based on mutational signatures helps to predict the outcomes as well as identify molecular subtypes of the disease. In gastric cancers, a signature found in several molecular subtypes was found to be predictive of survival in one type but not the others (Li et al. 2016), and signature analysis identified in esophageal cancer identified subgroups of tumors with different aetiology (Secrier et al. 2016). In B-cell lymphomas, analysis of clustered signatures yielded two signatures which could be used

to classify the tumours by their cell of origin (Alkodsi et al. 2019).

However, despite the potential of mutational signatures to improve the treatment of patients, several current limitations hold the medical oncology community from adopting this tool. More cancer data and confirmation experiments are needed to define the universal signatures which can serve as a reference, more clinical trials are required to confirm the utility and reliability of signatures as a response prediction indicator (Van Hoeck et al. 2019). Most of all, the mutational signature analysis currently lacks a unified and transparent methodology which would ensure the reproducibility and the best informative interpretability of the signatures under consideration.

1.2.7 Limitations of current mutational signature analyses

The majority of currently existing factorisation-based approaches suffer the drawback of having to choose the number of signatures. All of them employ different heuristics to identify the most suitable number of factors: quality of reconstruction, stability of signatures, biologically relevant restrictions on the number of signatures per patient and the number of mutations a signature should be contributing to a tumour.

The automatic relevance determination utilised by (Kasar et al. 2015) relies heavily on the choice of priors for the matrices of signatures and exposures, which may also not be representative of the underlying biology. Based on the NMF objective, a higher number of factors tend to provide a better reconstruction. If the algorithm tends to choose the strongest signals as factors, they will be stable and reproducible. Hence, even upon the integration of these heuristics into the signature analysis, there is still a high chance to derive an over-segmented set of signatures, mostly representative of the most mutated samples in the dataset. Consequently, a careful assessment of the biological feasibility of the extracted signals is essential.

Moreover, many mutational signature analyses focus on the distributions of single nucleotide variants (SNVs). However, there are many other types of damage that may be more representative of the mutational process. The challenge for analysing other types of damage is the need to find an adequate classification which would yield biologically feasible results.

Classification of copy-number changes by size and type turned out to be beneficial for detecting the HR deficiency and distinguishing between the deactivation of BRCA1, BRCA2 and other HR system defects (Nik-Zainal et al. 2016, Davies et al. 2017). Similar analysis including the size of homology at breakpoints and correlations with SNV signatures yielded signature based HRD determination in breast and ovarian cancers. A more sophisticated classification of the copy number variants by many characteristics allowed

to extract signatures highly predictive of the treatment outcome for high-grade ovarian cancers (Macintyre et al. 2018). Finally, a thorough classification of all structural variants detected in over 2,700 whole-genome sequenced cancers revealed a high contribution of replication-dependent mutagenesis to the SV spectra in cancer (Li et al. 2017b).

Analysis of signatures of dinucleotide variants and small indels confirmed the presence of several processes that were known to be involved in carcinogenesis, such as UV-induced CC>TT mutations, PAH-associated CC>AA changes, and a high amount of indels in repetitive regions occurring in MMR-deficient cancers (Alexandrov et al. 2018). The next step in this regard would be to find a way of combining all types of variants and analysing mutational signatures in the context of any mutational events.

Many previous studies have also suggested that mutations are not evenly distributed across the genome. Some mutational and repair processes are transcription- or replication-dependent, e.g. they could result from damage repaired by TC-NER, such as smoking-induced mutations (Pleasance et al. 2010), or they could be the result of polymerase errors (Reijns et al. 2015, Seplyarskiy et al. 2016). Hence the mutational signature may have a strand bias: they can be more or less pronounced in template versus coding strand, or in early versus late replicating regions (Morganella et al. 2016, Haradhvala et al. 2016, Tomkova et al. 2018). Some processes such as UV-light exposure have more specific genomic preferences: UV signature was shown to demonstrate different context specificity in gene promoters compared to the rest of the genome (Fredriksson et al. 2017).

Besides, certain mutational processes, such as somatic hypermutation or AID/APOBEC activity, can create distinct clusters of mutations (Supek and Lehner 2017, Roberts et al. 2012). Currently, strand biases or clustering of mutational signatures are investigated post hoc, after extracting the signatures from mutation counts. However, incorporating this information in the signature extraction has the potential to produce more biologically relevant signatures.

Overall, the field of mutational signature analysis currently suffers from two main drawbacks: the lack of generalised framework to assess performance which would allow choosing the best method and making it a golden standard, and a large pool of unexplored features of mutational processes that could refine the signature extraction and help to produce signatures with more biological meaning.

1.3 Aims of this thesis

The field of mutagenesis has been extensively studied over the last century and can nowadays provide an exceptionally detailed chemical and physical picture of individual events damaging the DNA, as well as their repair. The introduction of mutational sig-

natures allows us to take a broader look at the impact of DNA damaging agents on the genome. This scale of investigation provides a quantitative and qualitative understanding of the intensity and pathogenicity of genotoxic exposures and DNA repair deficiencies.

Analysis of mutational signatures in cancer can provide important information about the potential causes and vulnerabilities. About a half of mutational signatures encountered in cancer studies have an association with a mutagenic process, nearly two-thirds of them also have corresponding mutation spectra observed in model system experiments. Yet a large amount of factors and variability in mutational signature analysis remains unexplained, and requires further consideration.

Thus, the study I am describing in this thesis will contribute to these two aspects of the mutational signature analysis: identifying signatures of the factors potentially implicated in cancer and describing their mutational mechanisms, and explaining the variation in mutational spectra by considering the interplay between DNA damage and repair which may result in a dramatic change of mutational patterns in the genome. In this thesis, I am aiming to fulfil the following tasks:

- create a detailed catalogue of a wide range of DNA repair deficiencies and genotoxin exposures using a large mutagenesis screen conducted in *C. elegans*;
- provide mechanistic insights into the interaction of mutagens, DNA repair factors, and DNA;
- quantify the variability of experimental mutational signatures;
- compare the experimental mutational signatures to the ones observed in human cancers;
- describe the diversity of interactions between DNA damage and DNA repair;
- quantify contributions of different factors shaping mutational signatures in the experimental model system;
- explore the range of such interactions detectable in cancer.

In chapter 2 I will describe the model system, introduce the data and point out the most remarkable aspects which will be considered in greater detail in the following chapters. I will also introduce the main computational tools used for the analysis of the mutational spectra and distributions throughout the thesis.

Chapter 3 will present the analysis of mutational signatures and genomic properties of the mutations induced via DNA repair deficiencies. It will include an overview of mutation rates generated upon knockouts of different DNA repair factors, and a thorough

analysis of variants acquired upon translesion synthesis and homologous recombination deficiencies. Chapter 4 will focus on the signatures of mismatch repair deficiency and present a comparison between the signatures extracted from gastrointestinal cancers and *C. elegans* mutants. It will demonstrate the utility of model organism experiments in cancer research, where many interacting processes are active at the same time, and a fully unsupervised analysis is not sufficient to dissect the individual factors.

Chapter 5 will consider signatures of 12 genotoxins compared to their human counterparts. Combining deficiencies and exposures will also allow me to explore the diversity of DNA damage-repair interactions. I will show the most striking examples as well as review the frequency and magnitude of such effects in Chapter 6.

In Chapter 7, I will present an analysis of the DNA repair significance and its interactions with genotoxic exposures in cancer exomes. As a tumour is a complex and evolving structure, I will also present evolutionary considerations defining the scope of expected effects of the damage-repair interactions on cancer development and incidence.

Finally, in Chapter 8, I will summarise the findings and review the limitations of the study described in the previous Chapters. Additionally, I will outline the future perspectives and the directions of research that could further expand the understanding of the antagonism between the DNA damage and repair, as well as its implications for cancer and ageing.

Chapter 2

Experimental and computational methods to study mutagenesis in *C. elegans*

2.1 Introduction

The first chapter of this thesis gave an overview of the history and current state of knowledge of endogenous and exogenous mutagenesis and introduced mutational signatures in the context of analysing somatic mutations in human cancer. Based on this vast amount of knowledge, I will expand our understanding of mutational signatures, their translatability across species, and their behaviour under different genetic conditions.

Mutational signatures have become a useful tool for oncological investigations, providing the means to uncover the functional causes of tumour development. Some of the computationally extracted signatures were associated with the underlying processes, and a small fraction was confirmed experimentally. However, there are several questions which remain unclear:

1. How reliable are these associations? Are the relevant signatures directly linked to the associated factor?

2. What are the origins of the remaining signatures?

3. How consistent is the signal across different organisms, tissues, and genetic backgrounds?

To answer these questions, I will use a large experimental dataset and explore the effects of DNA repair deficiencies, genotoxic exposures, and their combination on the

genome-wide distribution of mutations. Further, I will compare the results obtained from *C. elegans* to the spectra observed in cancer, and develop the way to quantify the variability in mutational signatures in response to genetic background and other factors.

In this chapter, I will introduce the mutagenesis screen performed in *C. elegans* and describe the bioinformatics pipeline used to deliver the results from the sample stage to mutational spectra, creating the basis for studying experimental signatures of different factors contributing to mutagenesis.

Contributions

The experimental work on sample preparation was conducted at the University of Dundee by Bettina Meier and colleagues. DNA sequencing, alignment and basic variant calling for base substitutions and insertions/deletions were done by the Sanger Institute facilities. Subsequent bioinformatics analyses such as filtering, classification and analyses of mutational load and distributions as well as visualisations were performed by me.

All the data from this screen is available at the European Nucleotide Archive (Leinonen et al. 2010) under accession numbers ERP000975 and ERP004086, and the code used for these analyses is available under `http://github.com/nvolkova`. Different parts of these analyses were published in Meier et al. 2018, Volkova et al. 2019, or are currently under preparation to be submitted. The following chapters, which cover these analyses in more detail, will also provide a detailed description of the contribution in respective studies.

2.2 *C. elegans* as an experimental system for mutagenesis studies

In this study, we used *C. elegans* as a model organism to present a systematic screen of mutational signatures induced by genotoxins and DNA repair deficiencies, with the same setup as in Meier et al. 2014.

This organism is a suitable experimental model in many aspects: it is easy to manipulate, it has a well-annotated genome (Antoshechkin and Sternberg 2007) as well as a short turnover time with a lifespan of approximately 3 days (Hope 1999). *C. elegans* genome is small (approximately 100 Mbps) and has a high proportion of genome consisting of functional elements (about 30%, compared to just 2% in humans; Hillier et al. 2005). Importantly, DNA repair is conserved between human and *C. elegans*: 63% (75/118) of the core DNA repair genes in humans have close orthologs in *C. elegans* (Shaye and Greenwald 2011), and many other physiological processes are similar enough to use *C. elegans* as an emulator of human pathology (Kaletta and Hengartner 2006). Moreover,

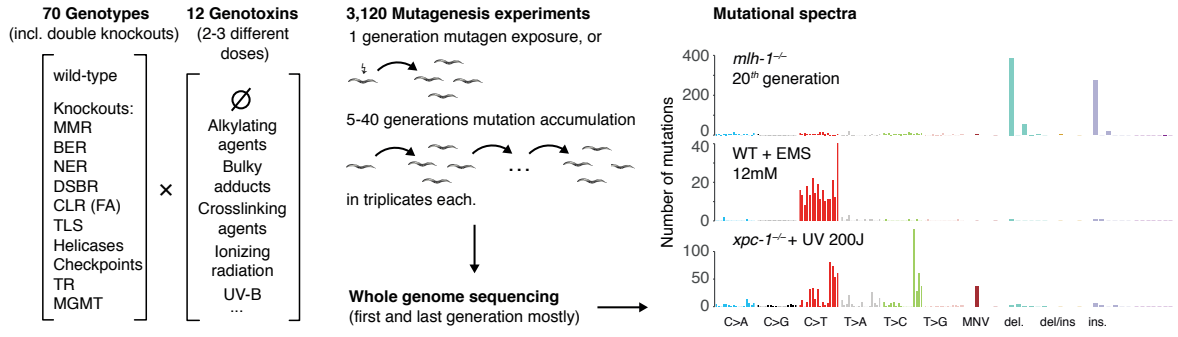
it has a low background mutation rate of about 1.8×10^{-10} mutations per site per cell division (Denver et al. 2009), which allows detecting even small changes in the rates of mutation acquisition.

In this screen, we can take advantage of the self-fertilising, hermaphroditic reproduction of *C. elegans*. About 15 cell divisions are required to pass from one generation to another, and both germ cells are coming from the same organism, thus creating a very stable pattern of mutation propagation across generations.

Moreover, the genome size of just about 100 Mbps makes DNA sequencing at 40x coverage cost effective at approximately £100 per sample. Consequently, given the presence of a well-established library of *C. elegans* DNA repair knockouts, we were able to perform a large screen with several thousand samples using the same material and computational resources as about 30 times smaller set of human cell line experiments.

2.2.1 Experimental design

The experimental design of the screen consisted of mutation accumulation and mutagen exposure experiments (Figure 2.1). In the mutation accumulation experiments, wild-type *C. elegans* and each of the 70 strains deficient for a particular DNA repair gene were propagated for several generations to measure the mutations accumulated during this period by subsequent whole-genome sequencing. In the mutagen exposure experiments, wild-type worms or worms with knockouts of different DNA repair components were exposed to different genotoxins, and their progeny was studied to analyse the range of germline mutations acquired during the self-fertilising stage of the parental generation.



wr

Figure 2.1: Experimental design of the study. *C. elegans* of different genetic backgrounds were propagated through several generations or exposed to metagenic agents. Their progeny was then sequenced to obtain the spectra of acquired mutations.

Mutation accumulation experiments

To mimic the mutation accumulation in somatic cells necessary for cancer development, wild-type *C. elegans* and *C. elegans* introduced with DNA repair deficiencies were grown through 20 or 40 generations, or for as many generations as possible (for the lines where the DNA repair deficiency was too deleterious). The experimental setup allowed propagating the clonal *C. elegans* lines, which in each generation passed through a single-cell bottleneck provided by the zygote. To filter out the mutations present in the line before the start of the experiment, the P0 (parental) or F1 (1st filial) generation were also sequenced. Each experiment was performed in triplicates when possible, and three samples were sequenced for each experiment.

List of genetic knockouts included genes representing all crucial DNA repair pathways: we consider 6 genes responsible for translesion synthesis (TLS), 13 backgrounds associated with single-strand break repair (SSBR) deficiency, 28 conditions associated with double-strand breaks repair (DSBR) deficiency, 6 helicases with different additional properties, 7 DNA damage checkpoint genes and 5 genes related to damage-caused apoptosis (see Table A.1, Appendix A). Due to the limitations discussed in Section 2.2, many of them did not show any particular patterns of mutagenesis. However, several most striking cases will be described in Chapter 3.

Mutagen exposure experiments

C. elegans were treated with DNA damaging agents at the late L4 and early adult stages to target both male and female germ cells. Resulting zygotes provide a single-cell bottleneck, where mutations are fixed before being clonally amplified during *C. elegans* development and passed on to the next generation in a Mendelian ratio. Samples of different genetic backgrounds, including wild-type, were treated with various genotoxins in triplicates. The maximal exposure dose was chosen such that it would generate as many mutations as possible without inducing severe mortality in the progeny (see Chapter 5 for more detail).

Panel of genotoxins used for mutagen exposure experiments consisted of substances and exposures employed in cancer treatments: alkylating agents (dimethyl sulfate (DMS), ethyl methanesulfonate (EMS), methyl methanesulfonate (MMS)), replication stalling drugs (hydroxyurea (HU)), irradiations (gamma-irradiation (IR), X-rays (Xray) and simulated UV-B irradiation (UV)), agent creating bulky adducts (aflatoxin-B1, aristolochic acid), and crosslinking agents (cisplatin, mechlorethamine and mitomycin C). The genotoxic agents, along with their signatures and mutagenesis features, are described in detail in Chapter 5.

In total, 915 experiments with 12 genotoxins and 70 different genetic conditions were performed. DNA was extracted from the parent worm and the last filial generation of these experiments, including 8 control experiments propagating wild-type *C. elegans*. Samples obtained from the experiments were subjected to next-generation sequencing using Illumina HiSeq short-read sequencing for the first 2027 samples, and 10x Genomics short-read protocol for the rest of the samples.

2.2.2 Pre-processing of sequencing data

Raw sequencing data was aligned to the WBcel235 assembly of the N2 Bristol strain reference genome using BWA (Li and Durbin 2009). Variant calling using a dedicated normal sample was performed separately for single nucleotide variants, small indels and large structural variants. The results have undergone a thorough filtering procedure to exclude technical errors, sequencing artefacts, caller mistakes and germline variance.

In brief, individual SNVs and medium-size indels obtained using CaVEMan (Nik-Zainal et al. 2012a) and PINDEL (Ye et al. 2009) variant callers, respectively, were subjected to the following filters:

- Coverage control: total coverage of the variant site in both the test and control samples should not exceed 150 reads or recede 15 reads;
- Absence of reads reporting the variant in the reference sample;
- VAF threshold: at least 20% of reads covering a site of interest in the test sample support the variant;
- Variant coverage: at least 5 reads support the variant in the test sample;
- PCR error control: at least one read in the test sample reports the variant in each direction;
- Indel filter for SNVs: should be no indel called at the same position (relevant for homopolymer junctions);
- Repetitive region artefact control: if the variant falls into a repetitive region, the regions should not be longer than 18 repeats;

Additionally, we implemented the deduplication procedure such that any variants repeated in unrelated samples (i.e. those which are not descendants of each other) were removed.

After filtering, SNVs for every sample were classified into 96 categories based on the trinucleotide context and the type of base change. Multiple substitutions which were found at adjacent sites in the same sample are classified as dinucleotide or multi-nucleotide

variants, if their VAF is similar (difference less than 5%). Indels were further classified based on the type (deletions, insertions, and complex indels - or deletions-insertions (DI)) and size (1 bp, 2-5 bp, 5-50 bp, 50-400 bp). Small insertions and deletions are further classified based on the local context: if the indel happened in repetitive sequence or not.

The structural variants were called using DELLY (Rausch et al. 2012) which extracts breakpoints based on paired end and split read mapping. Filtering for the raw variant calls included quality control ('PASS' filter reflective of the mapping quality, and at least 10 reads reporting the variant in the test sample), absence of the variant in control, and removal of artefacts/irrelevant events by removing all the events encountered in unrelated samples, or shared by a set of low-generation (0^{th} or 1^{st} generation) samples of the same genotype. Variants in telomeric regions were further removed due to complexity of resolving the repetitive structure or telomeric regions.

Structural variants were reported in the form of pairs of breakpoints, which were further classified in line with Li et al. 2017b into the following categories: tandem duplications (TD), deletions (DEL), inversions (INV), complex events (COMPLEX) - those with more than two pairs of breakpoints, intrachromosomal translocations (INTCHR), interchromosomal translocations (TRSL), and foldbacks (FOLDBACK) (when one inversion-like breakpoint is present, i.e. polymerase is turning around and reversing the DNA without turning back again).

The final distribution of variants among different classes after the filtering procedures is shown in Figure 2.2.

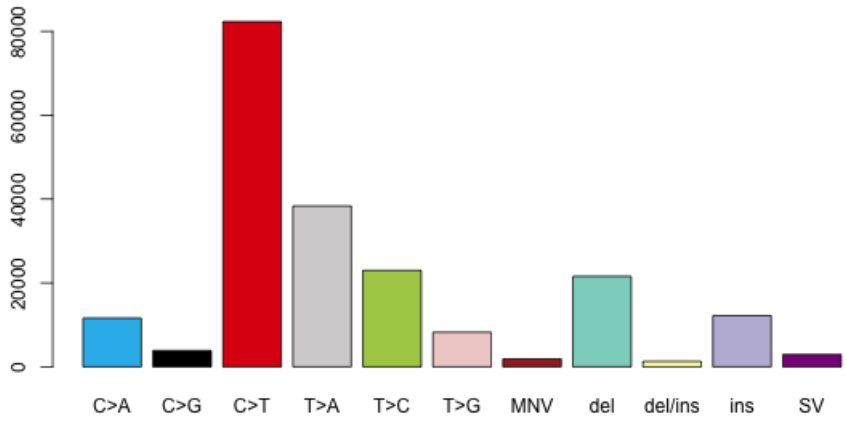


Figure 2.2: Overall distribution of mutations across all samples after filtering. MNV: multi-nucleotide variants, del - deletion, ins - insertion, SV - structural variants.

After performing quality control and variant calling, mutational spectra for each worm

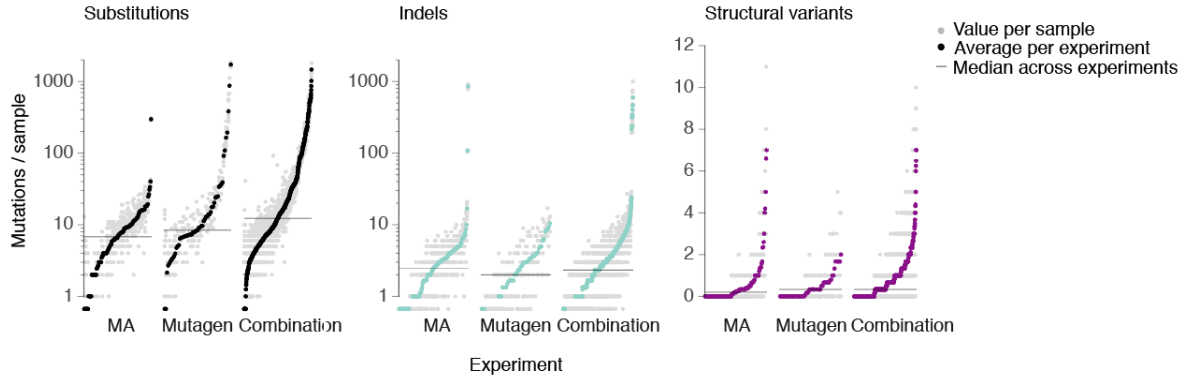


Figure 2.3: Number of observed mutations per replicate (grey dots) and experiment (average across replicates, colored dot) for base substitutions, indels and structural variants across different experimental types. Horizontal lines indicate the medians across experiments. Corresponds to Figure 1B from Volkova et al. 2019.

were obtained, including their base substitutions, insertions and deletions, and large-scale structural variants. In total, we called 190, 388 variants, of which 152, 237 were single-base changes, 1, 125 multi-nucleotide substitutions, 31, 702 indels and 2, 662 structural variants (Figure 2.3).

Variant classification

Filtered variants were classified into 119 mutation classes: 96 single-base substitutions classified by change and 5' and 3' base context, di- and multi-nucleotide variants (MNV), 6 types of deletions of different length and context, 2 types of complex indels differing in size, 6 types of insertions and 7 classes of structural variants.

Given the complementarity of base pairing, 6 possible base changes, namely C>T, C>A, C>G and T>A, T>C and T>G can be defined. We included the local context of these mutations in the form of their trinucleotide context describing the base upstream and downstream from the mutation, leading to a spectrum of 96 types of single-base substitutions. Several multi-nucleotide substitutions were also observed in the data. Upon checking the variant allele frequency of the base substitutions involving adjacent sites, we classified them as di- or multi-nucleotide variants (DNVs or MNVs) if the difference in VAF was less than 0.01.

Based on local characteristics, all the indels found in the data were classified by size: single-base events, 2 to 5 base pairs, 5 to 50 bps, and medium-sized indels of 50 to 400 bps. Small events involving one or 2-5 bps were also classified based on their local context: the sequence upstream and downstream from the indel was tested, and indels happening in repetitive regions (homopolymers, di-, tri-, tetra- or pentapolymers) were classified in

a separate category.

Structural variants were called in the form of breakpoints, which were clustered by proximity and classified based on the set of breakpoints within the cluster into tandem duplications, deletions, inversions, complex events, intrachromosomal translocations, interchromosomal events, and strand foldbacks.

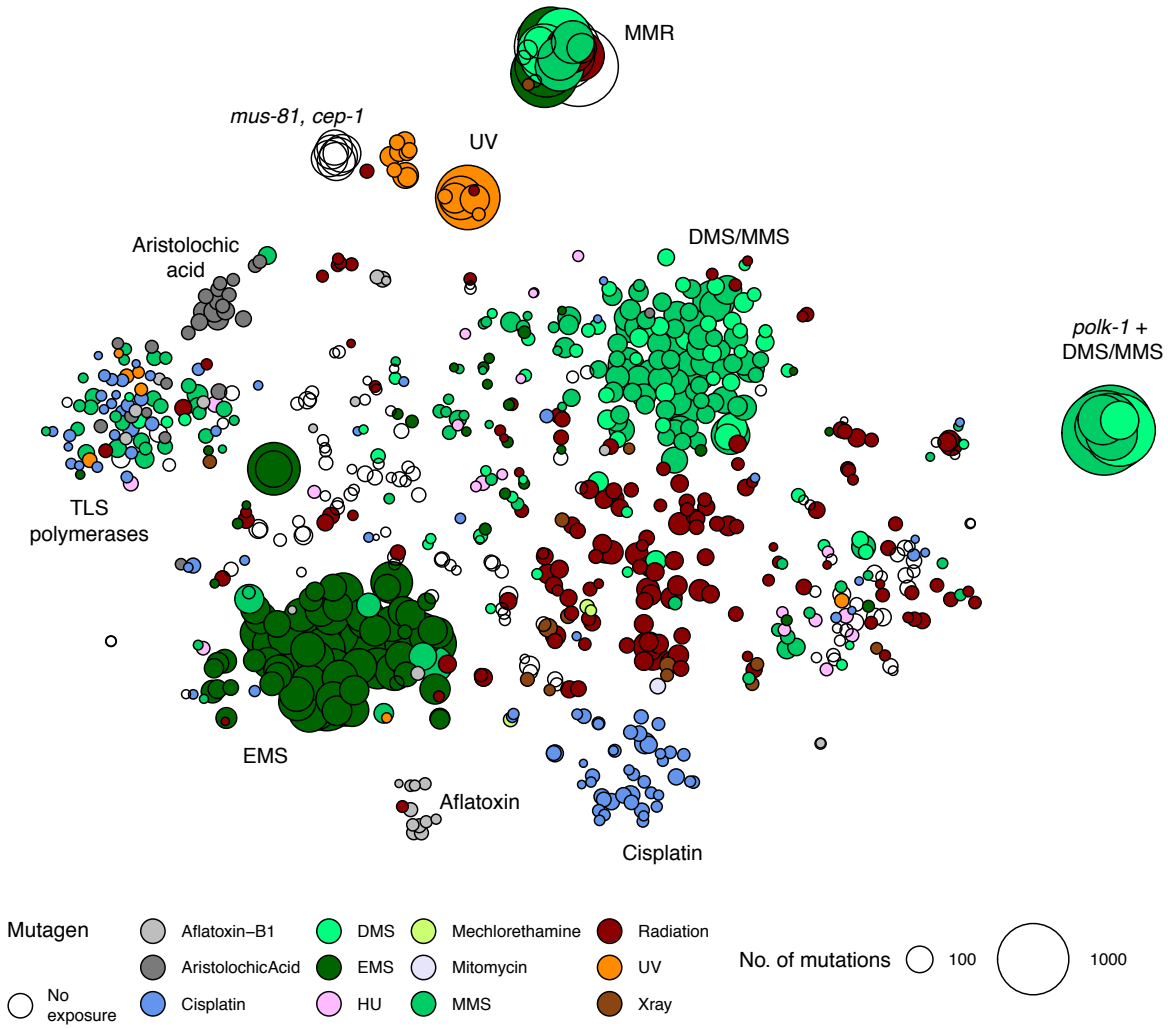


Figure 2.4: Similarity map of all the experiments in the screen. Each circle corresponds to a single experiment (average across three replicates), color reflects the mutagenic exposure, and the size reflects the number of mutations.

2.2.3 Overview of the data

Thus, we have obtained 3,120 119-long mutational spectra from 915 experiments with different genetic backgrounds and exposure doses. To obtain a panoramic overview of the mutation spectra and their mutual similarity, we created a t-SNE plot of the samples

where the distance between samples reflects how similar their spectra are to each other (Figure 2.4). The similarity between the spectra was calculated via cosine similarity:

$$\text{similarity}(\boldsymbol{S_1}, \boldsymbol{S_2}) = \frac{(< \boldsymbol{S_1}, \boldsymbol{S_1} >)}{(||\boldsymbol{S_1}|| \cdot ||\boldsymbol{S_2}||))},$$

where $< \boldsymbol{S_1}, \boldsymbol{S_2} >$ is a scalar product of signature vectors, and $||\boldsymbol{S_1}|| = \sum_{k=1}^{119} S_1^k$.

Clearly, the mutational spectrum of each sample was mostly defined by mutagenic exposure. Samples with different genetic background exposed to the same mutagen clustered together, indicating that the genetic background usually has only a moderate influence on the spectrum. However, a number of exceptions to this trend were visible, which will be further investigated in Chapters 3 and 6. Mismatch repair mutants (MMR) were separated from the rest of the samples, and no exposure could overwhelm the MMR deficiency spectra (Figure 2.4, top). Translesion synthesis (TLS) polymerase knockouts produced spectra similar to each other regardless of the mutagenic exposures (Figure 2.4, left).

2.3 Extracting mutational signatures from experimental data

To utilise the controlled nature of the experiments, we used a mathematical model describing the contributions of different factors to the mutational spectra of each sample. We aimed to model the full mutational profile of a sample j as a function of its genetic background and exposures. The natural assumption for mutation counts is a Poisson generative model. In order to preserve the additive relationship between different factors, we initially employed an additive Poisson model:

$$\boldsymbol{Y}_j = \{Y_{i,j}\}_{i=1}^{m}$$

$$\boldsymbol{Y}_j \sim \text{Poisson}(\boldsymbol{\lambda}_j),$$

$$\text{E}[\boldsymbol{Y}_j] = \boldsymbol{\lambda}_j = g_j \left(\boldsymbol{G}_j \times \mathbf{S}_G^T\right) + d_j \left(\boldsymbol{M}_j \times \mathbf{S}_M^T\right),$$

where $Y_{i,j}$ was the number of mutations of type i in sample j (and $\boldsymbol{Y}_j$ - a row vector of mutation counts for samples j), g_j reflected the number of generations that sample j was propagated for (under the assumption that all mutations acquired in each generation are passed to the next one), $\mathbf{S}_G \in \text{Mat}_{m\times p}\left(\mathbb{R}^+ \cup \{0\}\right)$ – signatures of DNA repair deficiencies, $\mathbf{S}_M \in \text{Mat}_{m\times r}\left(\mathbb{R}^+ \cup \{0\}\right)$ – signatures of mutagens, $d_j \in \left(\mathbb{R}^+ \cup \{0\}\right)$ – the dose of the mutagen for sample j, and the row vectors $\boldsymbol{G}_j \in \{0,1\}^p$ and $\boldsymbol{M}_j \in \{0,1\}^r$ were indicator vectors containing the information about the DNA repair knockouts and mutagen

exposures in sample j (given that there are overall p knockouts and r mutagens).

If we concatenated the genetic and mutagenic signatures in one matrix, this model could be reformulated to make similar to the NMF framework described in equation 1.3 with a fixed exposure matrix E. Hence, we had a computationally efficient way of simultaneously estimating all parameters in this multivariate regression task.

However, a closer look at the residuals within replicates upon fitting a Poisson model suggested the presence of overdispersion: residual deviance was significantly greater than the residual degree of freedom (res.dev = 51697.44 whereas df = 42615, chi-squared p-value < 0.001), therefore correct assessment of parameter variance required a model which could account for the additional variance.

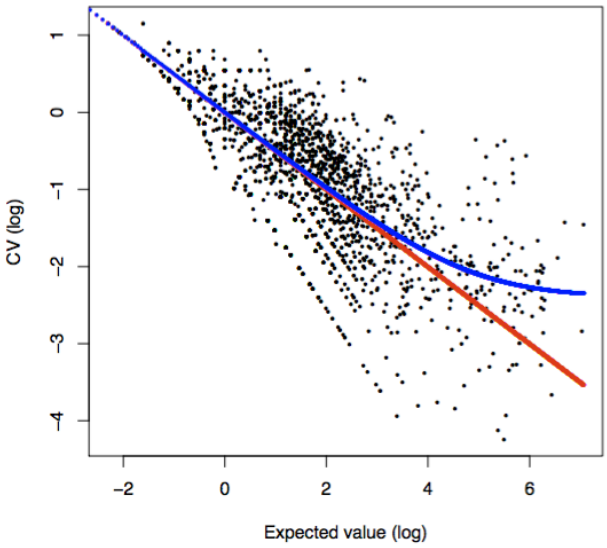


Figure 2.5: Per-experiment per-mutation type coefficient of variation (CV) to mean relation (log scale). Red line corresponds to Poisson distribution, blue line corresponds to maximum-likelihood fit using negative binomial distribution with the same dispersion parameter for all experiments.

Hence, we decided to allow for additional variation in the data by adopting a negative binomial distribution for the $n \times m$ matrix $\mathbf{Y}$ of observed counts:

$$\mathbf{Y} \sim \text{NegativeBinomial}\left(\boldsymbol{\mu}, \phi\right), \tag{2.1}$$

where the overdispersion parameter $\phi = 100$ was chosen empirically based on the estimates of overdispersion in the dataset (Figure 2.5).

2.3.1 Selecting the appropriate model

The naive model we utilised as the first approach assumed that the signatures of genotoxins should be consistent across genetic backgrounds, and that the resulting profile

should be a linear combination of the signatures of genotoxins and DNA repair deficiencies. If this model was correct, it would be possible to fit the samples combining DNA repair deficiency and genotoxin exposure using the signatures extracted from mutation accumulation and wild-type exposure experiments. However, careful inspection of the outliers demonstrated that the prediction was wrong across multiple experiments, suggesting that additional interaction factors should be included to account for changes in genotoxin-induced mutation rates compared to the wild-type experiments, as well as for the complete change of mutational spectrum.

In order to be able to inspect and compare the degree of interaction between experiments, we represented the mutational contribution of genotoxins in sample j as

$$d_j \left(M_j \times S_M^T\right) \exp\left(W_j * S_I^T\right),$$

where $W_j \in 0, 1^s$ is an indicator vector reflecting the presence or absence of genotype-genotoxin combination, and $S_I^T \in \mathrm{Mat}_{m \times s}(\mathbb{R})$ is a matrix of log-interaction effects.

Due to a higher mutation burden in interaction samples compared to mutation accumulation samples, a straightforward estimation of parameters was prone to disbalance between the values of the signatures of genetic backgrounds and interaction effects due to lower costs of wrongly predicting the outcomes for mutation accumulation samples (Figure 2.6). Hence, we have applied restrictive priors on the matrices of signatures and multiplicative interaction effects to ensure that mutational signatures are mainly estimated from the mutation accumulation and wild-type exposure experiments, while the experiments combining different factors will only inform the interaction matrix.

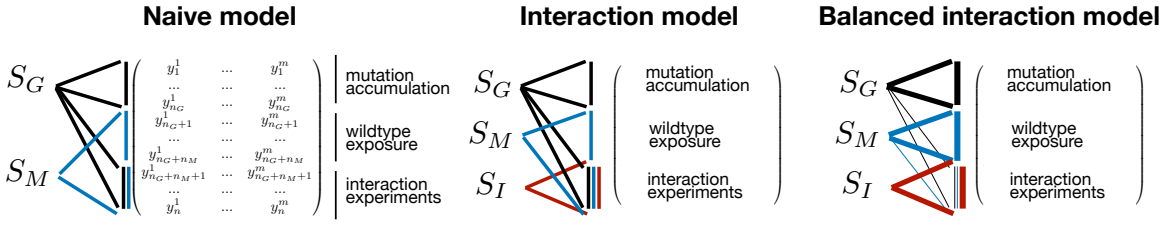
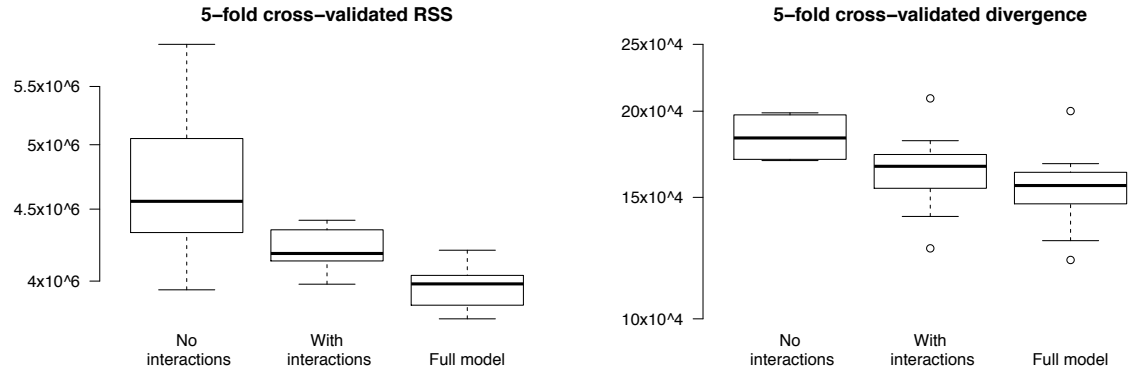


Figure 2.6: Schematic depiction of different models for extracting signatures and interaction effects from experimental data

Moreover, a hierarchical structure also simplifies identification of significant effects by degrading the task of signature comparison down to assessing whether the interaction effect is different from zero. In addition, it allows to control multiple confounding factors including misannotation of initial generation number for exposure experiments, and variation in precise mutagen dosage between different batches of samples, as will be described below (Section 2.3.2) when introducing the parameters of the full model.

To assess the ability of the model to correctly predict the profiles for new data, we performed 5-fold cross-validation (Mosteller and Tukey 1968) and estimated the reconstruction error and Kullback-Leibler divergence between the predicted and observed values for the test set (Figure 2.7). Introduction of interactions provides a great improvement over the naive model, while adding the parameters controlling confounding factors helps to perform better on the relevant subgroups of samples.



(a) (b)

Figure 2.7: Residual sum of squares (RSS) (a) and average KL-divergence (b) estimated by 100 runs of 5-fold cross-validation in three models: without interactions, with genotype-genotoxin interactions, and full hierarchical model.

2.3.2 Hierarchical model for dissecting the contributions of different factors

To quantify how much the signatures of genotoxins can change across different genetic backgrounds, we created a hierarchical framework which allows small alterations to the signature spectra for the samples with a combination of a knockout and exposure. The framework we suggest is a hierarchical Bayesian model schematically depicted in Figure 2.8.

As before, let $\mathbf{Y} \in \mathrm{Mat}_{n\times m}(\mathbb{N}\cup\{0\})$ be the matrix of mutation counts, which describes the mutational spectra over m types of mutations for n samples. Let there also be p different genetic backgrounds, r mutagenic exposures, and s combinations of DNA repair knockouts and genotoxic exposures. Additionally, let $\mathbf{G} \in \mathrm{Mat}_{n\times p}(\{0,1\})$ be an indicator matrix of genotypes across samples, $\mathbf{M} \in \mathrm{Mat}_{n\times r}(\{0,1\})$ – indicator matrix of exposures, $\mathbf{I} \in \mathrm{Mat}_{n\times s}(\{0,1\})$ – an indicator matrix of interactions, and $\mathbf{J} \in \mathrm{Mat}_{n\times(s+r)}(\{0,1\})$ – an indicator of experiments with interaction or any genotoxic exposure.

There are two main contributions to the expected value of the count matrix $\mathbf{Y}$: genetic contribution $\boldsymbol{\mu}_G$ and mutagenic contribution $\boldsymbol{\mu}_M$, defined by the genotype and

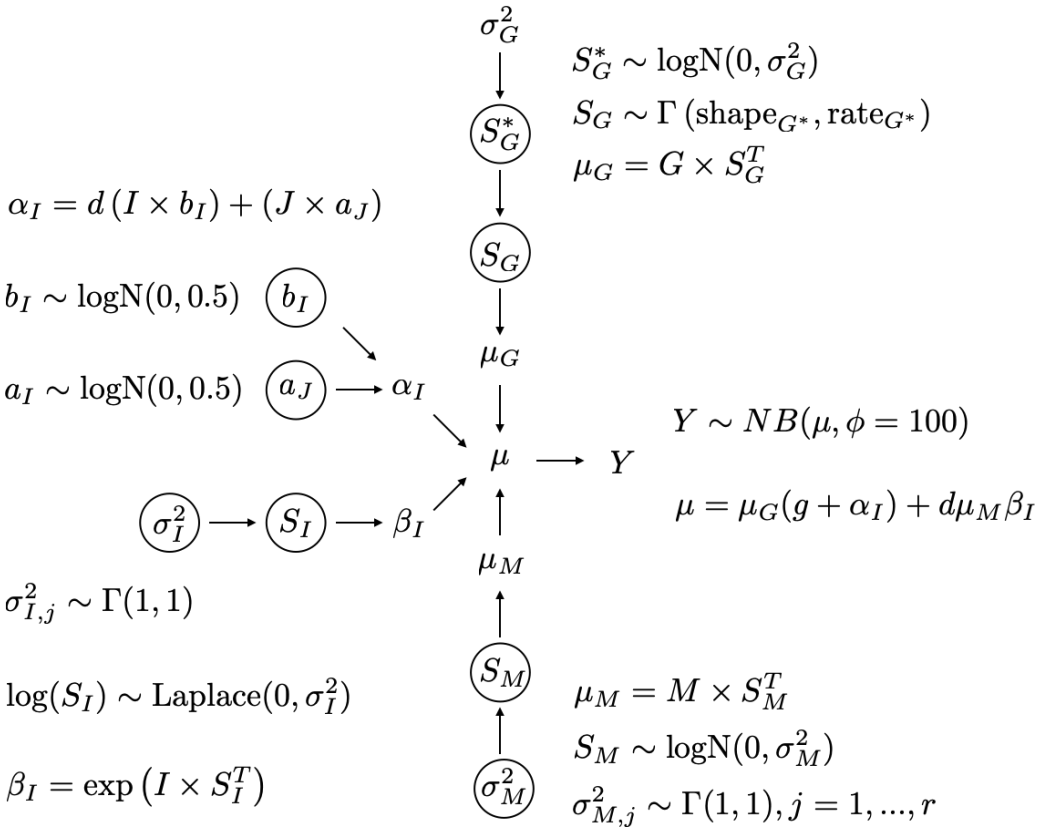


Figure 2.8: Graphical representation of the interaction model. Round circles denote the stochastic components. Y is the observed set of counts.

mutagen indicator matrices $\mathbf{G}$ and $\mathbf{M}$, respectively, and the signature matrices $\mathbf{S}_G \in \text{Mat}_{m\times p}\left(\mathbb{R}_+ \cup \{0\}\right)$ and $\mathbf{S}_M \in \text{Mat}_{m\times r}\left(\mathbb{R}_+ \cup \{0\}\right)$ that we aim to determine:

$$\boldsymbol{\mu}_G = \mathbf{G} \times \mathbf{S}_G^T, \boldsymbol{\mu}_M = \mathbf{M} \times \mathbf{S}_M^T. \tag{2.2}$$

Interactions for $\boldsymbol{\mu}_M$

Since the mutagens were shown to be the major defining factor of the mutational spectra (Figure 2.4), we first focused on the interaction effects altering the appearance of $\mathbf{S}_M$. To account for these interaction effects, we introduced a multiplicative term $\boldsymbol{\beta}_I \in \text{Mat}_{n\times m}(\mathbb{R})$ which describes the change of mutagen signatures in the samples with a combination of DNA repair knockout and mutagen exposure:

$$\boldsymbol{\beta}_I = \exp\left(\mathbf{I} \times \mathbf{S}_I^T\right).$$

This term is composed of $\mathbf{I}$, an indicator matrix of s interactions and their presence in n samples, and $\mathbf{S}_I, \mathbf{S}_I \in \text{Mat}_{m\times s}(\mathbb{R})$ - a matrix of changes which DNA repair deficiencies cause to the mutational spectra of the respective mutagen for each combination. As $\mathbf{S}_I$

can be both positive and negative, both negative and positive effects can be accounted for. To allow for more flexibility in extreme values, we chose a Laplace prior on $\mathbf{S}_I$: for each interaction j, $\boldsymbol{S}_I^{(j)} \sim \text{Laplace}(0, \sigma_{I,j}^2)$, and all $\sigma_{I,j}^2 \sim \Gamma(1,1)$ iid.

Essentially, each entry of β_I represents a fold-change in a particular type of mutations from the mutational spectrum of a mutagen caused by the interaction with the genetic background. Thus, the total contribution of the mutagens in the model is described as $\boldsymbol{d}(\boldsymbol{\mu}_M \cdot \boldsymbol{\beta}_I)$, where $\boldsymbol{d}$ is a vector of mutagen doses for all n samples, and the operator $\cdot$ means element-wise multiplication.

Interactions for $\boldsymbol{\mu}_G$

The genetic contribution, however, also carried several sources of variability. The mutations contributed by the DNA repair deficiencies in the absence of any mutagenic exposure are likely to result from endogenous damage, which would be consistent in both the samples with and without additional exposures. Hence, we do not expect any negative effects on the mutational spectra generated by DNA repair defects. Consequently, we introduced an additive adjustment tern $\boldsymbol{\alpha}_I \in (\mathbb{R}_+ \cup \{0\})$.

First, the samples used in genotoxic exposure experiments may have slightly diverged from the first filial generation F1. Hence, the term $\boldsymbol{\mu}_G$ (Equation 2.2) would not correctly describe the genetic contribution part in these samples. As the precise generation number for these samples was not available, we described these additional generations in each experiment with exposure by the vector $\boldsymbol{a}_J \in (\mathbb{R}_+ \cup \{0\})^{s+r}$, which means $\mathbf{J} \times \boldsymbol{a}_J$ additional generations per sample. Offsets $\boldsymbol{a}_J$ were modelled as log-normally distributed, $\boldsymbol{a}_J \sim \text{logN}(0, 0.5)$.

Moreover, we noticed that there are cases when the exposure causes a dose-dependent amplification of the spectrum of the DNA repair deficiency rather than an original or altered mutagen signatures, such as UV or MMS exposure of TLS polymerase knockouts. This type of interaction was usually weak compared to the average amount of mutations induced by a genotoxin, and could not be captured by $\boldsymbol{\beta}_I$ effects. Hence, we added a term $\boldsymbol{b}_I \in (\mathbb{R}_+ \cup \{0\})^s$ which described the amplification of genotype spectrum caused by an average dose of a mutagen in each of the s interactions. Across all samples, it would mean additional $\boldsymbol{\mu}_G\,(d(I \times b_I))$ mutations of the mutagen-free spectrum. Rates $\boldsymbol{b}_I$ were also modelled as log-normal distributed random variables, $\boldsymbol{b}_I \sim \text{logN}(0, 0.5)$.

Thus, the overall contribution of the genetic background can be described as:

$$\boldsymbol{\mu}_G\,(g + (\mathbf{J} \times \boldsymbol{a}_J) + d(\mathbf{I} \times \boldsymbol{b}_I)) = \boldsymbol{\mu}_G\,(g + \boldsymbol{\alpha}_I)\,.$$

Identifying signatures and interaction spectra

The model altogether aims to estimate the signature matrices $\mathbf{S}_G$ and $\mathbf{S}_M$, the interaction coefficients $\mathbf{S}_I$ and the additional variation vectors $\boldsymbol{a_J}$ and $\boldsymbol{b_I}$ based on the following structure:

$$\begin{aligned} \mathrm{E}[\mathbf{Y}] = \quad & \boldsymbol{\mu}_G(g + \boldsymbol{\alpha}_I) + d(\boldsymbol{\mu}_M \cdot \boldsymbol{\beta}_I) = \\ = \quad & (\mathbf{G} \times \mathbf{S}_G^T)\,(g + (\mathbf{J} \times \boldsymbol{a}_J) + d(\mathbf{I} \times \boldsymbol{b}_I)) + \\ & + d\left((\mathbf{M} \times \mathbf{S}_M^T) \cdot \exp\left(\mathbf{I} \times \mathbf{S}_I^T\right)\right). \end{aligned} \tag{2.3}$$

The total number of parameters in this model (Equation 2.3) is comparable to the number of observations, hence parameter fitting can lead to uncertain values due to unidentifiability issues. To avoid it, we exploit the fact that we have a separate set of mutation accumulation experiments without these additional variability factors, which allows for shrinking the priors of $\mathbf{S}_G$ by introducing a two-step fitting procedure. We first estimate the posterior distributions for the signatures of DNA repair deficiencies, which will, in turn, be used as priors for genotype contributions in interaction experiments. This way, we both ensure the biological feasibility of these factors and incorporate the signals from interaction experiments to better refine them.

Thus, in the first stage, the mutation counts $\mathbf{Y}^*$ for each of the mutation accumulation samples are modeled by a negative binomial distribution

$$\mathbf{Y}^* \sim \mathrm{NegativeBinomial}(\boldsymbol{\mu}^*, \phi = 100),$$

with an expectation

$$\boldsymbol{\mu}^* = g^* \cdot (\mathbf{G}^* \times \mathbf{S}_{G^*}^T).$$

Here the asterisk means that the respective entity was restricted to the samples from mutation accumulation experiments. A log-normal prior $\mathbf{S}_{G^*} \sim \mathrm{logN}(0, \sigma_G^2)$ iid with a scalar variance σ_G^2 was used for quantifying S_G^*. In total, we used 451 samples from mutation accumulation experiments with generation number higher than 1, and obtained signatures for 70 genotypes (the rest of the knockouts did not have mutation accumulation experiments).

In the main step, the prior for the signatures of genetic knockouts $\mathbf{S}_G$ was defined as $\mathbf{S}_G \sim \Gamma(\mathbf{shape}_{G^*}, \mathrm{rate}_{G^*})$, where $\mathbf{shape}_{G^*}$ and rate_{G^*} were fitted to the posterior draws for $\mathbf{S}_{G^*}$.

Signatures of genotoxins contributed more mutations on average and exhibited less variability. They were fitted simultaneously with the rest of the parameters using a individual lognormal prior for each mutagen, $\boldsymbol{S}_M^{(j)} \sim \mathrm{logN}\left(0, \sigma_{M,j}^2\right)$, where the variances

$\boldsymbol{\sigma}_M^2 = \left(\sigma_{M,1}^2, ..., \sigma_{M,r}^2\right)$ had prior distributions $\sigma_{M,j}^2 \sim \Gamma(1,1)$ iid.

Parameter estimation using Hamiltonian Monte-Carlo

Defined as described above, the posterior distributions of different parameter groups are intractable. This can be dealt with by using Markov Chain Monte-Carlo methods which aim to estimate the parameter of interest by randomly sampling from a probability distribution, which is achieved by constructing a Markov Chain that has the desired distribution (in this case, intractable posterior) as its equilibrium state. This way, we can obtain a sample from the posterior distribution to calculate point estimates as well as the variability of the parameters. Given the high dimensionality, we used one of the most efficient sampling methods - Hamiltonian Monte Carlo sampling which is based on sampling the derivatives of the target density function to generate efficient transitions that span the target posterior distribution (Betancourt and Girolami 2015, Neal and Others 2011).

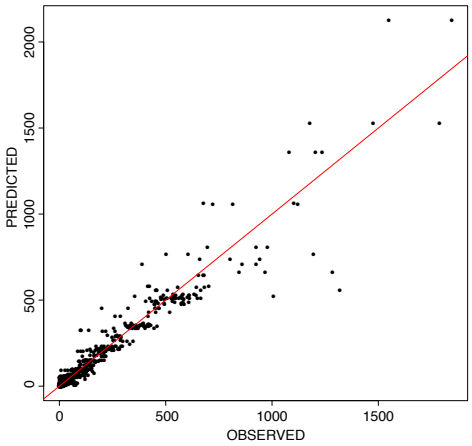


Figure 2.9: Observed and predicted mutation counts across all samples.

The model was specified using R-package "greta" (Golding 2018), and the posteriors of S_G, S_M, S_I, a_J, b_I and hyperparameters σ_M^2, σ_I^2 were estimated using Hamiltonian Monte Carlo sampling. We used 2000 steps for warmup and 5000 steps over 4 chains to ensure convergence. The point estimates for the parameters were taken as means of the samples across all chains. These estimates yielded a good agreement between the observed and predicted values of mutation counts (Figure 2.9).

2.3.3 Model for the simultaneous extraction of signatures and interaction effects for human data

We will also define a model for *de novo* signature extraction in the presence of additional factors that can alter the appearance of one of the signatures in a subset of samples. Assume there are N samples, and their mutational spectra consisting of M types of mutations are stored in a matrix $\mathbf{Y} \in \mathrm{Mat}_{N\times M}(\mathbf{N} \cup \{0\})$. Let $\mathbf{X} \in \mathrm{Mat}_{N\times K}(\{0,1\})$ be

an indicator matrix of the K factors that can alter the spectrum of a mutational signature, where each column is a binary vector describing the presence of a factor across the samples.

In a general mutation signature analysis framework described in Section 1.2, one seeks to represent a matrix of mutation counts $\mathbf{Y}$ as a product of an exposure matrix $\mathbf{E}$ and a signature matrix $\mathbf{S}$ under some assumptions about the nature of the noise (Equation 1.1). We will again assume a negative binomial generative model:

$$\mathbf{Y} \sim \text{NegativeBinomial}(\boldsymbol{\mu}, \phi),$$

$$\boldsymbol{\mu} = \mathbf{E} \times \mathbf{S}^T.$$

For the analysis of human data, we will use $\phi = 50$ chosen based on variability estimates across all cancer samples. Assume there are P signatures, $\mathbf{S} \in \text{Mat}_{M \times P}(\mathbb{R} \cup \{0\})$. To quantify how much a signature $\boldsymbol{S}^{(P)}$ changes in the samples with a particular factor, we will decompose the expectation as follows:

$$\mu = \mathbf{E}_{-P} \times \mathbf{S}^{-(P)} + (\mathbf{E}_P \times \mathbf{S}^{(P)}) \cdot \boldsymbol{\beta}_F. \tag{2.4}$$

The effect $\boldsymbol{\beta}_F$ is a matrix which defines how the signature $\boldsymbol{S}^{(K)}$ looks in each sample depending on its composition of factors. $\boldsymbol{\beta}_F$ can be represented as

$$\boldsymbol{\beta}_F = \exp\left(\mathbf{X} \times \mathbf{S}_X^T\right),$$

where $\mathbf{S}_X \in \text{Mat}_{M \times K}(\mathbf{R})$ is a matrix of spectra of the interaction effects per factor. We assumed a uniform prior for $\mathbf{E}$, $\mathbf{E} \sim \text{Unif}(0, R)$ where R is the maximal number of mutations per sample in the dataset, and a normal prior for $\mathbf{S}_X$, $\mathbf{S}_X \sim \text{N}(0, 0.5)$. This model can use any sort of binary information to infer its effects on the signature.

The posterior distributions for the signatures and the interaction effects were estimated using Hamiltonian Monte Carlo sampling (Neal and Others 2011) via R package 'greta' (Golding 2018). All models were run in 4 chains up to 1000 or 2000 warm-up and 1000 post-warm-up samples to ensure convergence. We run 4 chains of sampling and claimed an effect being real if it was consistently assigned to the same signature. The final number of signatures was selected based on the convergence, similarity between signatures and feasibility of effect assignment, as the model tends to fluctuate or duplicate most variable signals when the chosen number of dimensions is too high.

The efficacy of the method was demonstrated using a simulated dataset with 100 patients with at most 10000 mutations, 96 substitution types, 3 signatures (we took COSMIC signatures 1, 17 and 4) and 1 factor affecting the last signature in 20% of

samples. In 100 trials we observed the reconstruction error and average similarity between reconstructed signatures being the lowest and the similarity to the original signatures being the highest for $S = 3$, the real number of signatures (Figure 2.10). The error in effect estimation did not differ between $S = 2, 3, 4$, but was generally low confirming that the model is capable of simultaneously extracting correct mutational signatures and estimating the effects of additional factors on one of the signatures.

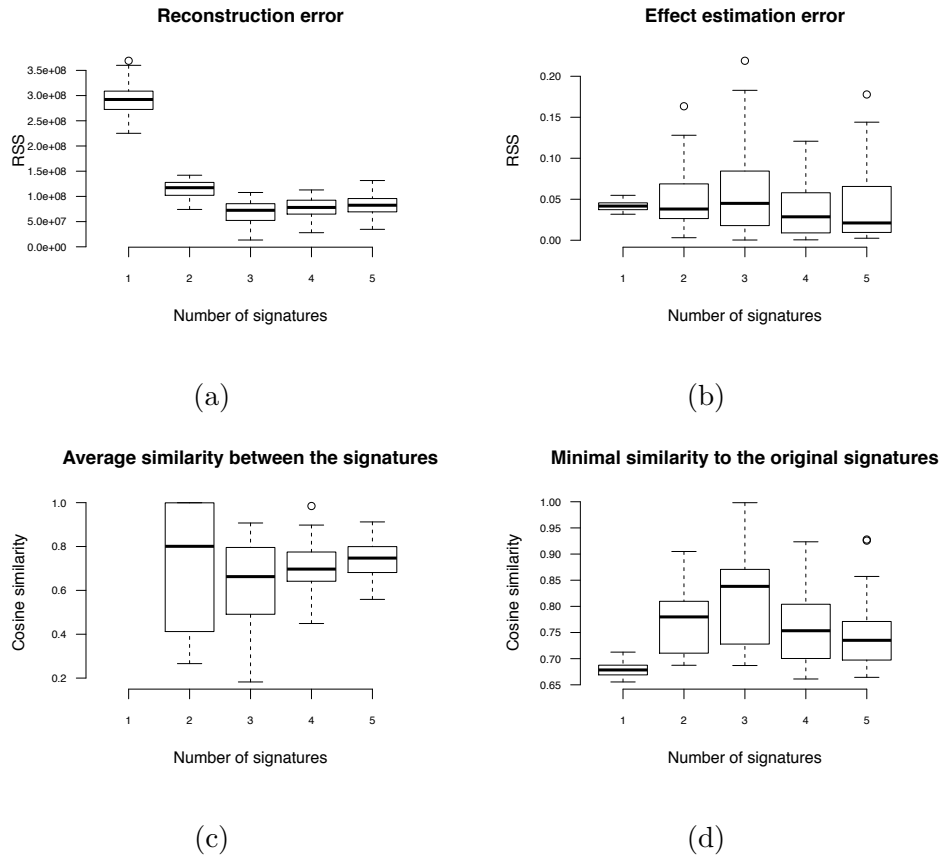


Figure 2.10: Reconstruction error (a), effect recovery error (b), average similarity between extracted signatures (c), and the minimal similarity to the original signatures (d) for different number of signatures in the simulated dataset with 3 underlying signatures with 1 factor.

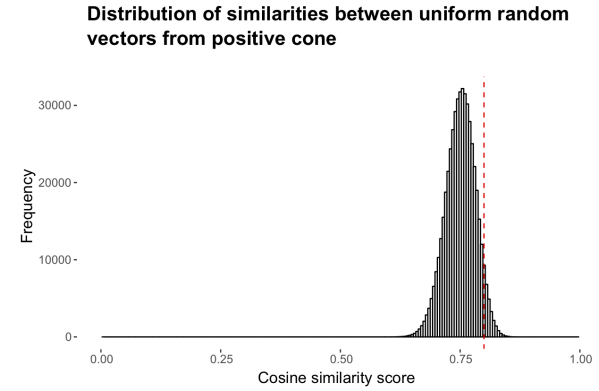
2.4 Mutational signature comparison

2.4.1 Assessment of cosine similarity score as a measure of similarity of mutational spectra

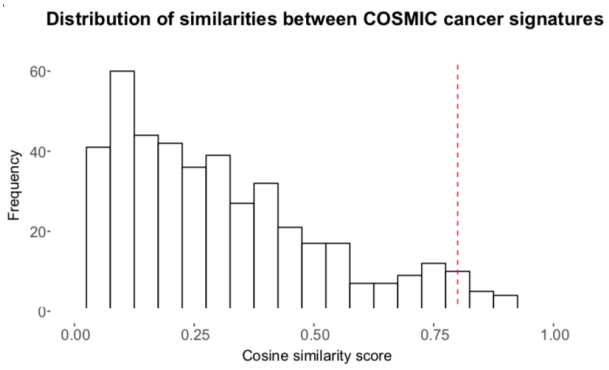
Cosine similarity is currently the most common measure of similarity between signatures. If the two signatures are identical, the similarity between them is 1. However, it is

not clear when to consider signatures dissimilar.

To find a threshold for 'high' similarity, we assessed the similarity distribution between random uniformly generated 'mutational profile' vectors from the positive cone (Figure 2.11a). Its 95% quantile falls in 0.80, which we used further on as the cutoff under which we can not call two signatures similar. Additionally, analysis of similarities within the COSMIC cancer signature set (Alexandrov et al. 2013b) showed that only 3.4% of pairs have a cosine similarity above this threshold (Figure 2.11b).



(a)



(b)

Figure 2.11: (a) Cosine similarity distribution for random vectors drawn uniformly from 0 to y_{max} (individual for every draw). The 0.95-quantile of this distribution is 0.80 indicated by a red dashed line. (b) Frequency of cosine similarity scores for all pairwise signature comparisons across the 30 COSMIC cancer signatures. 3% of signature pairs exhibited a similarity higher than 0.80, a threshold indicated by a red dashed line. Corresponds to Supplementary Figure 6A from Meier et al. 2018.

2.4.2 Comparing mutational signature across specie

Many DNA repair genes are conserved between *C. elegans* and humans; however, performing a valid comparison between the mutational spectra and factor interactions in these two systems is complicated by several factors, which require adjustment.

A major difference between *C. elegans* and *Homo sapiens* is their trinucleotide composition. The base frequencies of the *C. elegans* genome were skewed towards A and T, especially in repetitive contexts ApApA and TpTpT, providing higher chances of T>N changes in TpTpN context (Figure 2.12). The GC content of the *C. elegans* genome is 36% (Sequencing Consortium* 1998) which puts it close to the human genome with a GC content of 41% (Piovesan et al. 2019), but far from the human exome that has 64% of GCs (Lelieveld et al. 2015). Most of the currently available cancer data is coming from whole-exome sequencing; hence, our main objective was to be able to compare the signature observed from *C. elegans* to those found in human exomes.

Thus, to provide a meaningful comparison between *C. elegans* and cancer-derived mutational signatures, the experimental signatures acquired from *C. elegans* were adjusted to the human exome (or genome, if relevant) trinucleotide frequencies. The probabilities for 96 base substitutions were multiplied by the ratio of respective trinucleotide counts observed in the human exome (hg19, the counts pre-calculated in Rosenthal et al. 2016) to those in the *C. elegans* reference genome (Figure 2.12).

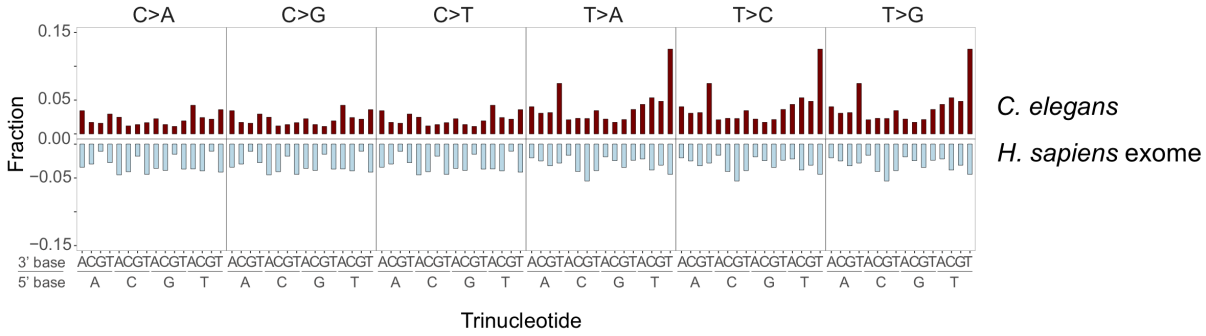


Figure 2.12: Trinucleotide context comparison between *C. elegans* genome and *H. sapiens* exome.

Apart from a 10-fold difference in size, the architecture of the genome differs substantially between *C. elegans*, which has about 27% of the genome contained in exons (Sequencing Consortium* 1998), and the human genome with only about 1% of bases coding for proteins (Gregory 2005). Hence, we can expect fewer mutations in worms, and some of the observed spectra may be biased by such factors as transcription strand-specific DNA damage or DNA repair. In addition, a complex metabolic activation of certain genotoxins can make their mutational footprint different in cells directly exposed to the agent or indirectly exposed to its metabolites. The details of the metabolic activation of the genotoxins used in the screen will be described in Chapter 5.

Lastly, one has to remember that cancer cells contain a historical record of mutagenic exposures which may have changed over time. Consequently, they may have more factors

contributing to the intensity of the observed mutational effects, with unknown order and duration, which creates much higher variability in the mutational spectra than can potentially be observed in *C. elegans* within the same experiment.

2.5 Measuring other genetic features

2.5.1 Relationship to transcription and replication directionality

To investigate the relationship between mutations in regions with different directions of transcription, we created a map of transcription directionality using a list of all transcribed genes along with the direction of transcription with respect to the reference (Lee et al. 2017) and classified every genetic region into 3 categories: as being transcribed in '+' (e.g. the coding strand being the reference strand) direction, being transcribed in '-' direction, or ambiguous (possibly transcribed in both directions). In total, 93% of coding regions (which compose 67% of the *C. elegans* genome) were assigned a transcription direction.

The directionality of replication was determined using the Okazaki fragment sequencing from Pourkarimi, Bellush, and Whitehouse 2016. The fractions of Okazaki fragment reads on the minus strand,

$$t_j^l = \frac{t_{j-}^l - t_{j+}^l}{t_{j-}^l + t_{j+}^l}, j = 1, ..., 6, l = 1, ..., 1002685$$

were calculated for 100-bp bins. The bins where $|\text{mean}_j\left(t_j^l\right)| > 2\text{sd}_j\left(t_j^l\right)$ were assigned a "+" (or called right-replicating) direction if $\text{mean}_j\left(t_j^l\right) > 0$, or "-" (left-replicating regions) if $\text{mean}_j\left(t_j^l\right) < 0$. In total, we inferred the direction of replication for 45% of the genome.

2.5.2 Analysis of clustered mutations

The clustering of mutations along the genome was assessed using the starting points of all the base substitutions and indels across the samples of the same genotype and generation. Clustered status was assigned based on a hidden Markov model, which predicts a series of M hidden states $H = \{H_m\}_{m=1}^{M}$, $H_m \in \{\text{clust}, \text{not}\}$ (being in a cluster or not) for all mutations within a sample based on the set of distances to the next mutation $D = \{D_m\}_{m=1}^{M-1}$, $D_m \in \mathbb{N}$ (the last mutation in each chromosome is assumed to be fixed in non-clustered state). The probability of a set of states given the observed distances would then be calculated as

$$P\left(H_{1:M}, D_{1:M}\right) = P(H_1)P\left(D_1|H_1\right)\prod_{m=2}^{M-1} P\left(H_m|H_{m-1}\right)P\left(D_m|H_m\right),$$

where the transition probabilities

$$P\left(H_m = \text{clust}|H_{m-1} = \text{not}\right) = 0.001,$$

$$P\left(H_m = \text{not}|H_{m-1} = \text{clust}\right) = 0.1$$

and starting probabilities $P\left(H_1 = \text{clust}\right) = 0$, $P\left(H_1 = \text{not}\right) = 1$.

Each of the distances D_m can be considered as the number of nucleotides without mutation before the mutation with index m. Then $D_m|H_m$ can be considered as the number of failures before the first success in a Bernoulli trial with a success probability that depends on the state. Hence, the distances D given the states can be described by geometric distribution. The density of mutations within a cluster was assumed to be at least one mutation per 100 bases, and the mutations outside the clusters were assumed to be uniformly distributed:

$$D_m|H_m = \text{clust} \sim \text{Geom}\left(p = 0.01\right),$$

$$D_m|H_m = \text{not} \sim \text{Geom}\left(p = \frac{1}{\text{mean}(D)}\right).$$

We used the Viterbi algorithm (Viterbi 1967) to infer the most likely set of states for each sample.

2.6 Discussion

In this chapter, I introduced the dataset describing several hundred mutagenesis experiments conducted on wild-type and DNA repair-deficient worms by mutation accumulation over generations as well as by direct mutagen exposure. Having a system with controlled exposures to different mutagenic factors allows us to describe the mutational footprints of each factor, compare their relative contributions, and study the interaction effects between different groups of factors. The primary overview of the data suggested that the mutagenic exposures have a higher impact on the mutational spectra in the samples with a combination of DNA repair defect and genotoxic exposure.

We proposed a negative binomial regression framework to extract the contributions of different mutagenic factors. This model can capture both the scale and the spectrum of interaction-caused changes. Additionally, we proposed an algorithm for simultaneous *de*

novo extraction of signatures and their mixing coefficients from cancer data along with the effects of additional variables on signature spectra. The applications of these frameworks to real data will be demonstrated in Chapters 6 and 7, respectively. Methodologically, this is the first analysis of mutagenesis experiments which considers interaction terms between genetic and mutagenic factors.

Some limitations of the experimental approach have to be acknowledged. High fidelity of replication, self-fertilising mode and a smaller non-coding fraction of the genome limit the number of mutations that may be observed before the *C. elegans* lineage becomes sterile or non-viable (Thompson et al. 2013). The call to overcome this negative selection also limits the scale of events which may be observed: huge copy number rearrangements or chromosomal alterations are less likely to be observed in mutation accumulation experiments which involve germline propagation. Conversely, in human somatic cell line expansion, acquired copy number changes are quite common (Abyzov et al. 2012, Mishra and Whetstine 2016).

Nevertheless, this screen represents the first experimental dataset of this scale, covering the majority of DNA repair gene knockouts and exposures to several mutagens with different mechanisms of DNA damage. Based on these data and inference models, we described the mutational signatures of mutation accumulation across 70 lines with different genetic backgrounds and quantified the mutagenic effects of 12 genotoxins in the wild-type and DNA repair-deficient conditions. These results will be structured in Chapters 3, 5 and 6.

Chapter 3

Experimental signatures of DNA repair deficiencies in *C. elegans*

3.1 Introduction

In the previous chapter, I introduced the mutagenesis screen and described the tools and types of analysis used to produce an in-depth characterisation of experimental mutational signatures.

DNA repair deficiencies have been long acknowledged as one of the main driving forces in cancer. Many genetic conditions, which predispose to cancer, stem from mono- or bi-allelic deficiency in different DNA repair pathways. Heterozygous defects in mismatch repair machinery lead to Lynch syndrome which is associated with high rates of colorectal cancer (Cancer Genome Atlas Network 2012), homologous recombination deficiency conferred via mutations in BRCA1/2 genes yields high breast and ovarian cancer risks (Antoniou et al. 2003), nucleotide excision repair defects can lead to xeroderma pigmentosum, a syndrome with a tremendously elevated rate of skin cancers (Bradford et al. 2011).

Deactivation of a DNA repair pathway does not necessarily require a mutation. It may be a consequence of hypermethylation, or alteration in the regulation of transcription, or alteration of the post-translational modification process of a core component. Therefore, sequencing of a relevant gene or transcription measurements are not always representative of the actual functionality of a DNA repair pathway. In such cases, identification of a genome-wide signature of incorrect repair via this pathway is a more reliable way of detecting a deficiency (Hollstein et al. 2017, Van Hoeck et al. 2019). However, our understanding of the mechanisms of DNA repair is not extensive enough to be able to link every DNA repair deficiency to a pattern of genomic alterations. Research from geneti-

cally manipulated model systems is therefore crucial to the understanding of mutational consequences of DNA repair-related gene defects.

In this chapter, I will describe the quantitative characteristics and prospective mechanistic insights of the mutations introduced by DNA repair deficiencies in the *C. elegans* genome.

Contributions

This work was conducted in collaboration with Bettina Meier and colleagues in Anton Gartner's group at the University of Dundee and Peter Campbell at the Sanger Institute. All of the findings from this chapter will be included as a part of a manuscript under preparation:

Meier, B., Volkova, N.V., Hong, Y., Wang, B., Gonzalez-Huici, V., Bertonlini, S., Boulton, S., Campbell, P.J., Gerstung, M. and Gartner, A. Systematic analysis of mutational spectra associated with DNA repair deficiency in *C. elegans* mutation accumulation lines.

For this manuscript, I performed variant calling and filtering, as well as visualisations, analysis of clustering and local genomic features, and calculation of mutation rates and mutational signatures of DNA repair deficiencies. BM performed the in-depth analysis of structural variant breakpoints.

3.2 Mutation types and rates in wild-type and DNA repair-deficient strains

To characterise the mutational signatures of DNA repair deficiencies in *C. elegans*, we subjected 559 samples to mutation accumulation over 5 to 40 generations. In total, the dataset contained 57 single and 10 double mutants, which covered most of the conserved DNA repair and DNA damage response pathways (Table A.1).

We did not observe overtly elevated mutation rates in strains defective for nucleotide excision repair (NER), base excision repair (BER), non-homologous DNA end-joining (NHEJ) and apoptosis, consistent with a high level of redundancy of DNA repair pathways. Changes in mutagenesis were observed upon knockouts of translesion synthesis polymerases, highly conserved genes involved in homologous recombination (HR), upon depletion of mismatch repair machinery, and in double knockouts of HR-related and apoptosis-related genes. All in all, this chapter will provide a global analysis of how various DNA repair and DNA damage response pathways help to protect genome stability from endogenous mutagenesis.

3.2.1 Estimating mutation rates in mutation accumulation experiments

Accumulation of mutations was achieved by passing the *C. elegans* lines through several generations, introducing a single-cell bottleneck at each generation passage. We assume that all mutations accumulated in a single cell are heterozygous; they have 25% chance of being lost after propagation, 50% chance of remaining heterozygous, and 25% chance of being fixed (becoming homozygous). Hence, in order to make the number of mutations directly proportional to the number of generations, we adjusted the number of generations as

$$\widetilde{N} = \lambda(N) = \sum_{i=1}^{N} \frac{1}{2^{i-1}} + \frac{1}{4} \sum_{i=1}^{N-1} \sum_{j=1}^{i} \frac{1}{2^{j-1}},$$

where $\widetilde{N}$ denotes the adjusted generation number, and λ is a function reflecting the fraction of all accumulated mutations observed after N generations.

We calculated the mutation rate as a number of mutations of a particular type accumulated in a wild-type or DNA repair-deficient *C. elegans* line over one generation. Previously, we calculated a mutation rate per base pair of $\sim 1.0 \times 10^{-9}$ mutations per cell division for wild-type *C. elegans* (Meier et al. 2014, Meier et al. 2018). This corresponds to about one mutation in the one hundred million base pair genome within two *C. elegans* generations. For comparison, in humans, analysing parent-child trios yielded nearly 30 mutations occurring from one generation to the next (Conrad et al. 2011).

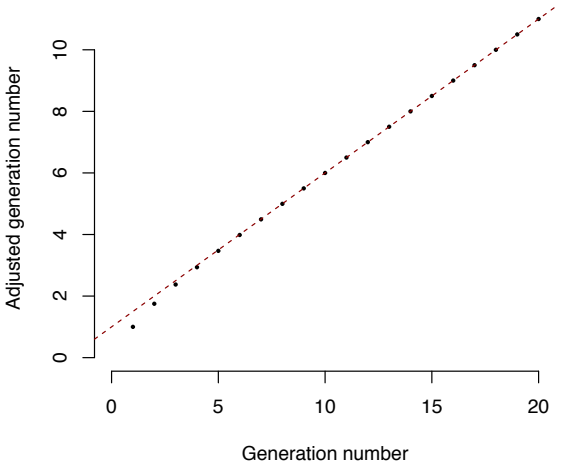


Figure 3.1: Adjusted generation number versus the real one. Dashed line represents the $y = x/2 + 1$ line.

Using the generation adjustment described above, we estimated the wild-type mutation rate as approximately one heterozygous mutation per genome per generation (0.95, $SD = 0.05$), which corresponds to about $N/2$ mutations after N generations as the function $\lambda(N)$ becomes indistinguishable from $N/2 + 1$ for any $N > 18$ (Figure 3.1).

3.2.2 Comparison of the mutation rates across genotypes

Different types of mutations, especially mutations of different scale, naturally have different occurrence rates, which makes a comparison of the total mutation rates across genetic backgrounds incorrect. Hence, we considered the rates of single-base substitutions, indels of up to 400 base pairs, and structural variants per generation separately. Median mutation rates across all genotypes were quite close to those in the wild-type: about 0.93 heterozygous substitutions (0.67, $SD = 0.05$ in wild-type), 0.33 indels (0.35, $SD = 0.03$ in wild-type) and 0.14 (0.12, $SD = 0.03$ in wild-type) structural variants per genome per generation.

Comparing the mutation rates across 66 DNA repair-deficient backgrounds to the wild-type (using a χ^2-test), we found that the substitutions rates for about half (53%) of all strains were not significantly different from the wild-type, and for another 25%, the change in mutation rate did not exceed 2-fold (Figure 3.2). Only 8 genotypes demonstrated substitution rates higher than two mutations per genome per generation: a knockout of double-strand break (DSB) repair regulator *rtel-1*, knockouts of stability maintenance complex elements *smc-5* and *smc-6*, knockouts of mismatch repair genes *pms-2* and *mlh-1*, and double knockouts *brc-1; cep-1*, *mus-81; cep-1* and *pole-4; pms-2*.

One-quarter of genotypes showed indel rates different from that in the wild-type, of which only 9 backgrounds (13%) exhibited at least a two-fold difference (Figure 3.2). Following our expectations, MMR-related knockouts *pms-2*, *mlh-1* and *pole-4; pms-2* demonstrated high rates of indels. Besides, the knockouts of TLS polymerases *polh-1* and *rev-3*, G-quadruplex resolving helicase *dog-1* and the double knockouts *brc-1; cep-1* and *mus-81; cep-1* also demonstrated elevated rates of indel acquisition.

In the structural variant rate comparison, only 20% of genotypes showed an SV rate different from the wild-type. In particular, the mutants defective in *helq-1*, *mus-81*, *slx-1*, *mus-81; cep-1*, *smc-5*, *rtel-1* and *smc-6* genes exhibited high SV rates (Figure 3.2). All of these genes are associated with DSB repair or resolution of crosslinks.

3.3 Experimental mutational signatures and genomic features

Upon analysing the mutational rates and signatures of DNA repair-deficient *C. elegans* lines, several pathways demonstrated alteration in mutation rates: mismatch repair, translesion synthesis mutants, lines defective in helicases which participate in crosslink repair, and homologous recombination deficient mutants (especially in combination with a knockout of an apoptosis-related gene). Below I will describe the most interesting spec-

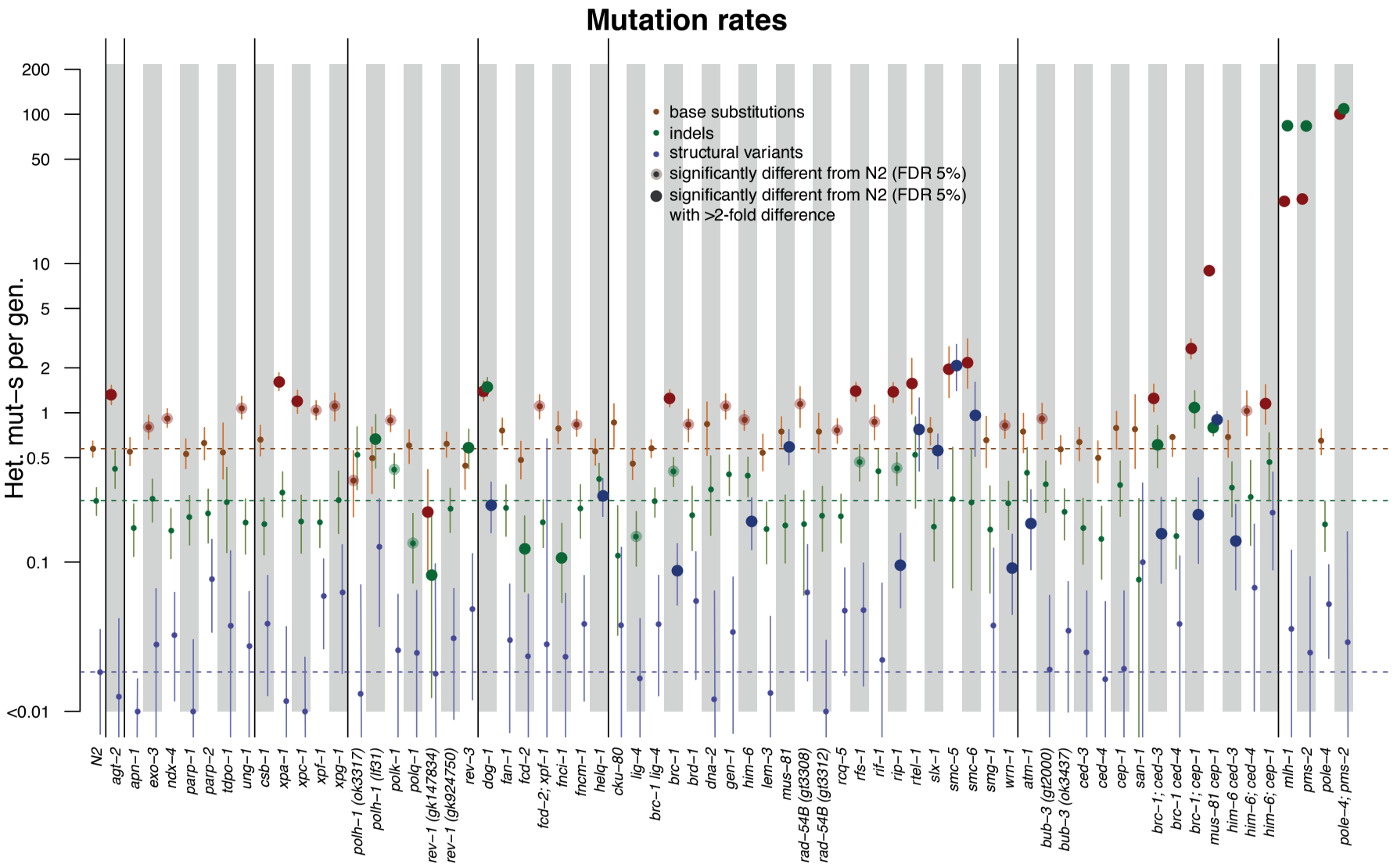


Figure 3.2: The rates of single-base substitutions (brown), indels (green) and structural variants (purple) in *C. elegans* across 68 genetic backgrounds with their 95% credible intervals.

tra and investigate additional features of mutations accumulated in these lines. A full list of all mutational signatures with indications of entries significantly different from the wild-type can be found in Appendix B.

3.3.1 Mismatch repair deficiency yields high rates of indels and single-base substitutions

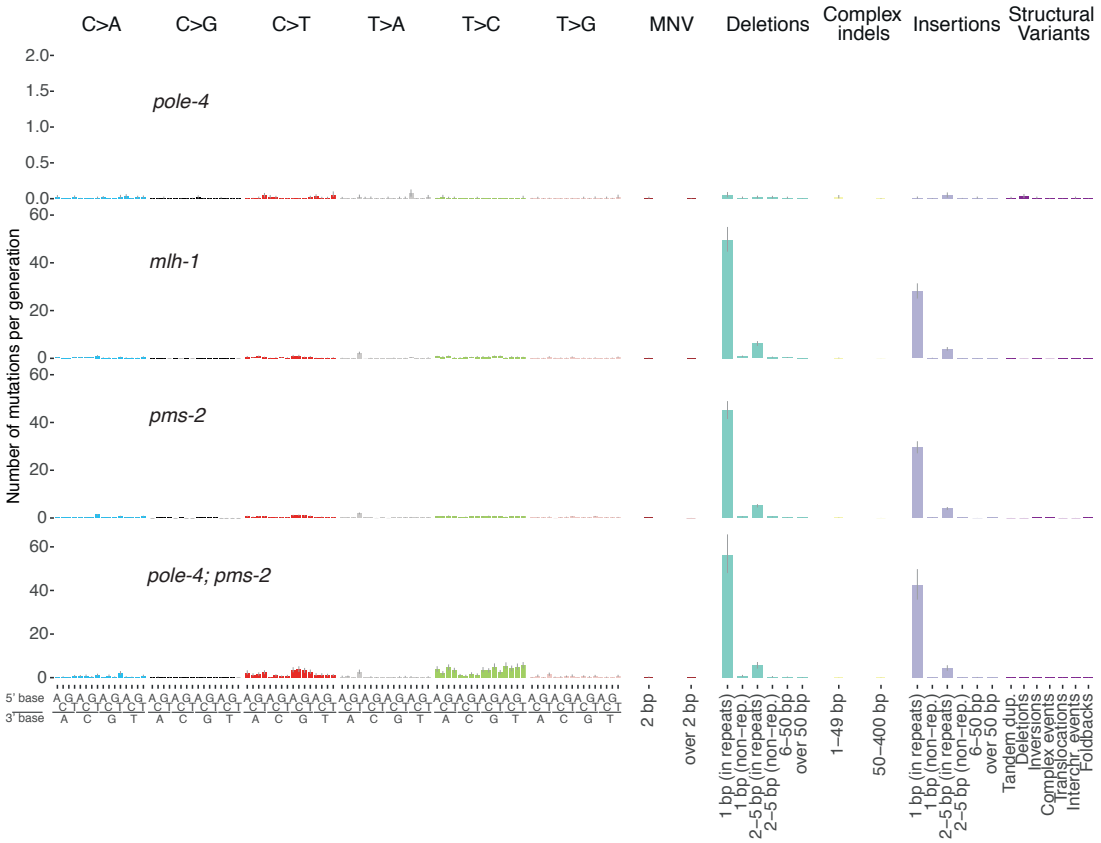


Figure 3.3: Experimental mutational signatures of *mlh-1*, *pms-2* and *pole-4; pms-2* deficiencies along with their 95% credible intervals. Distribution of mutations is expressed in numbers of mutations per generation.

Mismatch repair is one of the most crucial repair mechanisms, which dramatically decreases the replication error rate by correcting the mismatches left by the replicative polymerases ϵ and δ.

In contrast to the wild-type, *mlh-1* and *pms-2* mutants carried an average of 1174 and 1191 unique mutations, respectively, of which 288 and 309 were base substitutions and 886 and 882 indels. This corresponded to acquiring about 26 base substitutions and 83 indels per generation, or a total of 7.10×10^{-8} (95% CI: 6.86×10^{-8} to 7.33×10^{-8}) and 7.28×10^{-8} (95% CI: 7.10×10^{-8} to 7.48×10^{-8}) mutations per base pair and cell division

for *mlh-1* and *pms-2*, respectively.

Interestingly, a double knockout of *pms-2*, an MMR component, and *pole-4*, accessory subunit of the replicative polymerase epsilon, yielded a substantial increase in the base substitution rate. The double *pole-4; pms-2* mutants demonstrated a 2-fold increase in total mutation rate over the single MMR mutants, reaching 1.51×10^{-7} (95% CI: 1.45×10^{-7} to 1.56×10^{-7}) mutations per genome per division, leading to on average 101 base substitutions and 109 indels acquired per generation.

The main feature characterising the mutational spectra of MMR knockouts was a high amount of single-base indels in repetitive regions (Figure 3.3), indicating the polymerase slippage as the primary mechanism of mutation acquisition upon MMR deficiency. Presumably, additional defects in *pole-4* lead to an increase in the error rate of the polymerase, which manifests via higher base substitution rate. The mutational patterns and molecular mechanisms of MMR deficiency will be considered in more detail in the next chapter (Chapter 4).

3.3.2 Defective translesion synthesis yields medium-sized deletions

TLS polymerases are capable of replicating DNA across damaged bases. This synthesis can lead to an error-free or an error-prone lesion bypass. *C. elegans polh-1* (pol η) and *rev-3* (catalytic subunit of pol ζ) mutants displayed an increased level of deletions under 400 bps, especially those between 50 and 400 bps (Figure 3.4).

REV-1 protein serves as a scaffold protein which recruits other TLS polymerases to the damaged site. In the mutation rate comparison, *rev-1* knockout displayed a more than 3-fold decrease in mutagenesis compared to wild-type, which is consistent with having no TLS due to defective *rev-1* function (Figure 3.2). The mutational signature of this deficiency, however, was not different from the wild-type for any of the mutation classes. Interestingly, only *rev-1* (gk147834) demonstrated such behaviour, but not the other *rev-1* knockout, *rev-1* (gk924750).

Deletions associated with *polh-1* and *rev-3* deficiencies showed hallmarks of double-strand break repair via alternative end-joining, which is mediated via polymerase θ (POLQ). The involvement of POLQ is marked by microhomology at the break junctions. Additionally, there is evidence of polymerase slippage and re-priming at the junction, as reported previously (Roerink, Schendel, and Tijsterman 2014). All in all, our data suggest that POLH-1, REV-1 and REV-3 may prevent DNA breaks by reading across damage bases. The failure to do so apparently leads to small deletions. Our results on pol η/pol ζ defects and the evidence for repair by pol θ mirror the reports from the Tijesterman lab

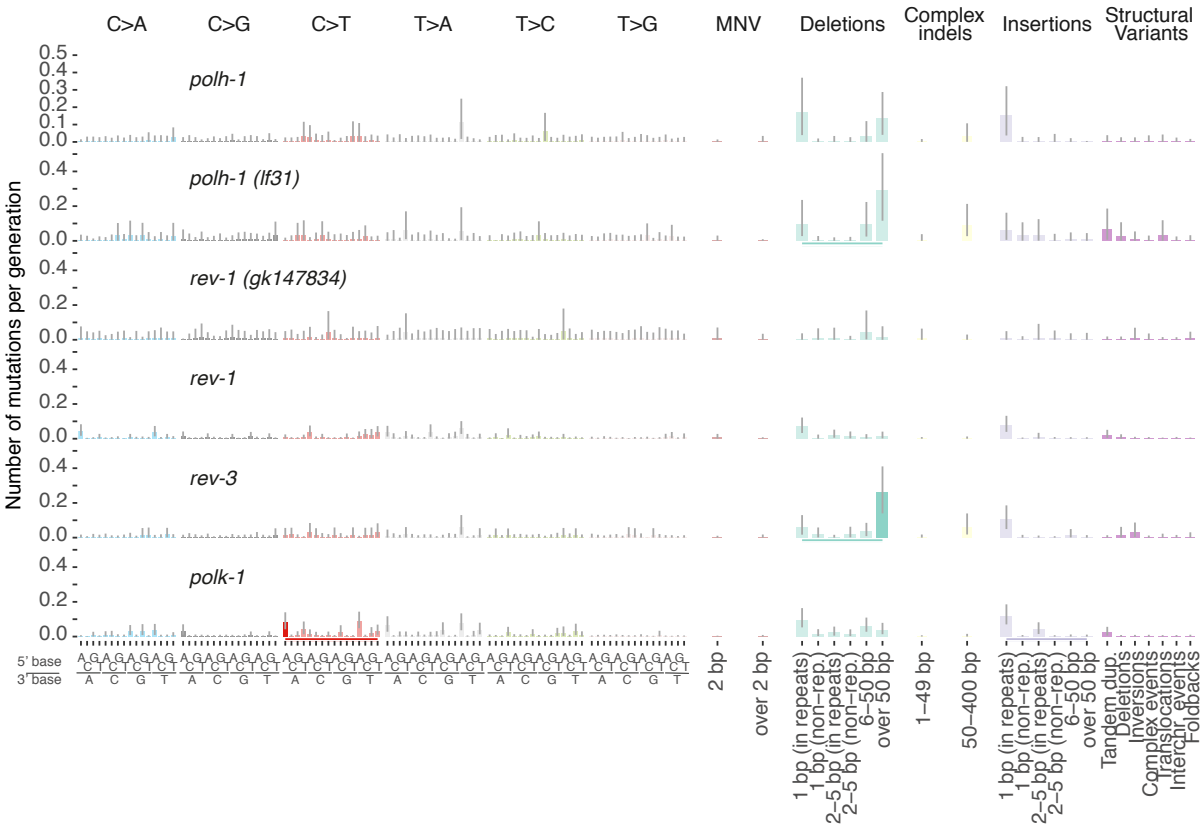


Figure 3.4: Mutational signatures of 4 TLS polymerase deficiencies, including two different knockouts for *polh-1* and *rev-1*. Brighter colors represent mutation contexts which are different from the wild-type. Lines underneath denote mutation classes where the total number of mutations within this class is different from the wild-type (FDR 10%).

(Roerink, Schendel, and Tijsterman 2014, Van Schendel et al. 2015).

3.3.3 Structural variation in DNA crosslink repair-deficient mutants

Repairing DNA interstrand crosslinks (ICL) usually involves the so-called Fanconi Anaemia (FA) pathway, which is required for sensing ICLs and for assembling various repair factors. Severing an ICL leads to a single-strand break (SSB) in one strand, which may be converted to a DSB and mended by HR repair if it persists until replication. Finally, TLS polymerases are also involved as they need to read across the remaining adduct on the other strand. In addition, the FA pathway is recruited to stalled replication forks to prevent genome instability upon DNA replication stress (Lachaud et al. 2016).

There are six ICL repair mutants considered in the study. FCD-2 (Fancd1 in mammalian cells) is a scaffold protein sensing the damage via ubiquitination upon ICL formation, and it is thought to be coordinating the assembly of the ICL repair complex.

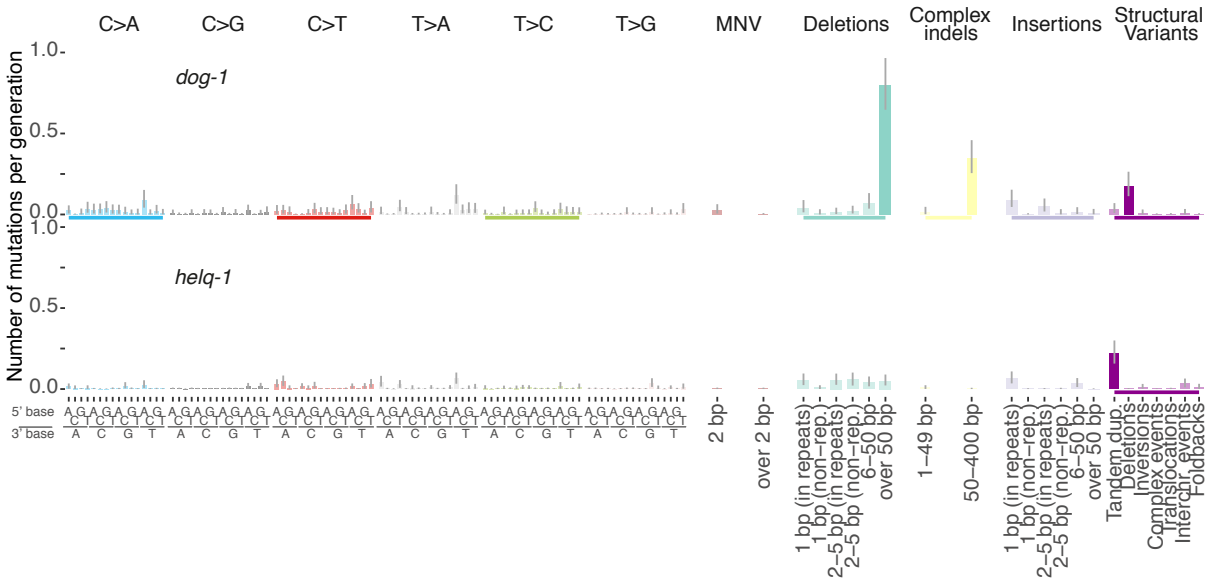


Figure 3.5: Experimental mutational signatures of helicase knockouts *dog-1* and *helq-1*. Brighter colors represent mutation contexts which are different from the wild-type. Lines underneath denote mutation classes where the total number of mutations within this class is different from the wild-type (FDR 10%).

FNCM-1 a helicase involved in DNA damage recognition. Together with FNCI-1, a component of an E3 ubiquitin ligase complex, they are needed for FCD-2 ubiquitination. FAN-1 is a structure-specific nuclease involved in ICL repair, but not epistatic to FCD-2 and FNCM-1. DOG-1 is the *C. elegans* ortholog of the mammalian FANCJ helicase implicated in the resolution of DNA tertiary structures formed by G-rich DNA stretches referred to as G-quadruplexes, and HEL-Q is a conserved helicase thought to act in parallel to FCD-2.

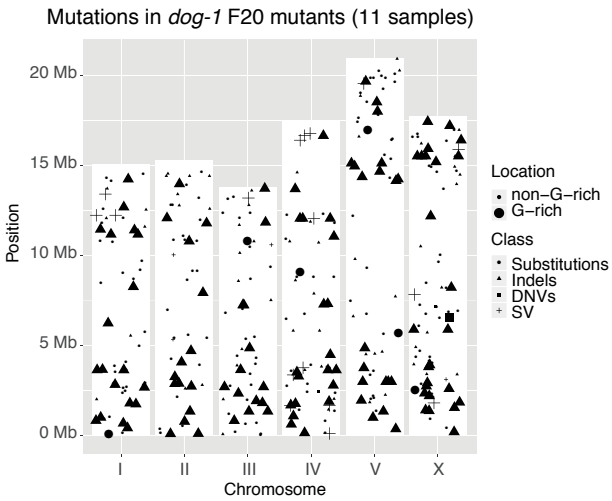


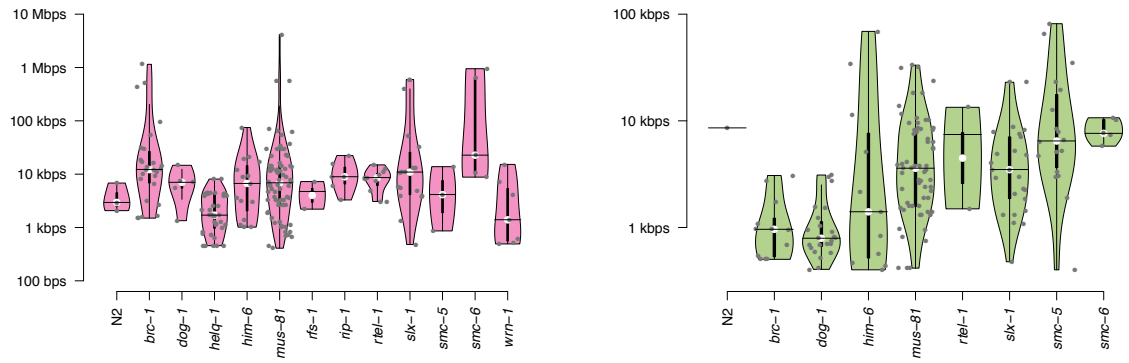
Figure 3.6: Distribution of mutations across 11 *dog-1* mutants. Large symbols correspond to the mutations which intersect with the motifs prone to form G-quadruplexes.

Analysing the mutational signatures present in the ICL repair mutants, we did not find an overt deviation between *fcd-2*, *fncm-1*, *fnci-1* and wild-type (Figure **??**). However, *dog-1* mutant strains showed a 3-fold increase in total mutation rates, and a distinct mutational pattern with a shift towards deletions and deletion/insertions (complex indels) in the range of 50-500 base pairs, as well as to larger SVs, predominantly deletions (Figure

3.5).

Across 11 *dog-1* deficient samples propagated for 20 generations, 81% of deletions (17 out of 21) longer than 400 bp were overlapping with regions prone to generate G-quadruplexes, as well as 78% (109 out of 139) of indels between 50 and 400 base-pair in length (Figure 3.6). The density of such G-rich motifs in *C. elegans* genome was shown to be about 0.89 per 1 kb (Marsico et al. 2019). Hence, our data suggest that the rate of G-quadruplex formation in *dog-1* deficient *C.elegans* is about 1 lesion per generation.

HELQ-1, a helicase previously named HEL-308, is thought to act in parallel to FancD2 in ICL repair. *helq-1* mutants showed a high number of tandem duplications, which ranged in size between 457 and 8089 bp with a median of 1270 bp (Figure 3.7a). The breakpoints of these tandem duplications tended to be enriched within sequences with inverted repeats. Of the 5 tandem duplications with breakpoints within an inverted repeat accumulated across 3 *helq-1* mutants after 40 generations, all occurred in right-replicating regions, which suggests that they may be arising in a replication-dependent manner due to secondary structures arising in single-stranded DNA.



(a) (b)

Figure 3.7: Distribution of (a) sizes of tandem duplications and (b) sizes of deletions accumulated in mutation accumulation experiments per genotype. Only genotypes with 2 or more deletions/duplications are shown.

3.3.4 Evidence of alternative DSBR under homologous recombination deficiency

Similar to mammalian cells, the repair of DSBs in *C. elegans* can occur via three routes: error-free homologous recombination (HR) repair, error-prone non-homologous end-joining (NHEJ), or alternative end-joining mostly comprised of the microhomology-mediated end-joining (MMEJ). Hence, the knockouts of either of the two systems should

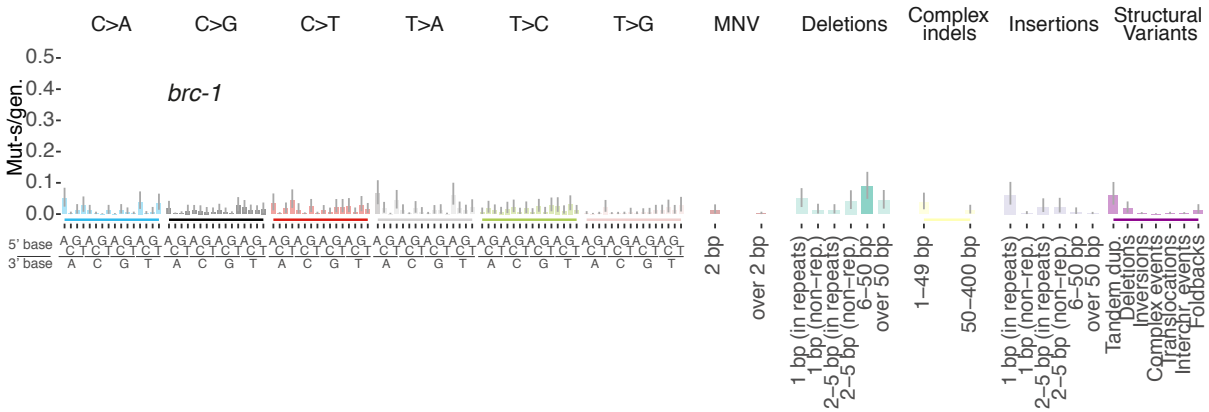


Figure 3.8: Experimental mutational signature of the *brc-1* knockout. Brighter colors represent mutation contexts which are different from the wild-type. Lines underneath denote mutation classes where the total number of mutations within this class is different from the wild-type (FDR 10%).

produce mutational spectra indicative of the activity of an alternative pathway.

Knockouts of NHEJ and MMEJ. We first analysed *cku-80* and *lig-4* mutants defective in NHEJ and *polq-1* mutants defective for MMEJ (Figure **??**). We did not see any biologically relevant difference in mutation rates for neither *lig-4* or *cku-80*, nor *polq-1* mutants, which confirms that homologous recombination is capable of maintaining genome integrity on its own, and does it in an error-free manner.

Deficiency in *BRCA-1*. In contrast, the *brc-1* mutants showed an increased number of mutations enriched with 5-50 bp pair deletions, as well as an increased SV rate (Figure 3.2). BRC-1 protein is involved in many stages of HRR and regulates several repair factors (Boulton et al. 2004), and an absence of this protein turns off the inter-sister recombination completely. Homologous recombination deficiency in *C. elegans* was associated with elevated rates of base substitutions, which resembled a flat profile of substitutions found in HR deficient human cancers (Polak et al. 2017, Riaz et al. 2017), deletions in the range of 5-50 bp, and larger tandem duplications (Figure 3.8).

The tandem duplications accumulated in *brc-1* mutants tended to have microhomology at the breakpoints, indicative of active MMEJ repair. Tandem duplications ranged in size between 1,601 bp and 500 kb, with a median of 11,913 bp, thus being longer than the duplications observed in *helq-1* mutants (Figure 3.7a). It also agrees with a pattern of structural rearrangements associated with BRCA-1 loss in breast and ovarian cancers (Nik-Zainal et al. 2016, Macintyre et al. 2018). In addition, *brc-1* mutants accumulated small deletions (Figure 3.7b) and base changes, which can also be a consequence of mutagenic NHEJ.

End-processing and sister chromatid invasion. Once a DNA double-strand break occurs, it has to be processed by a series of nucleases to generate single-stranded

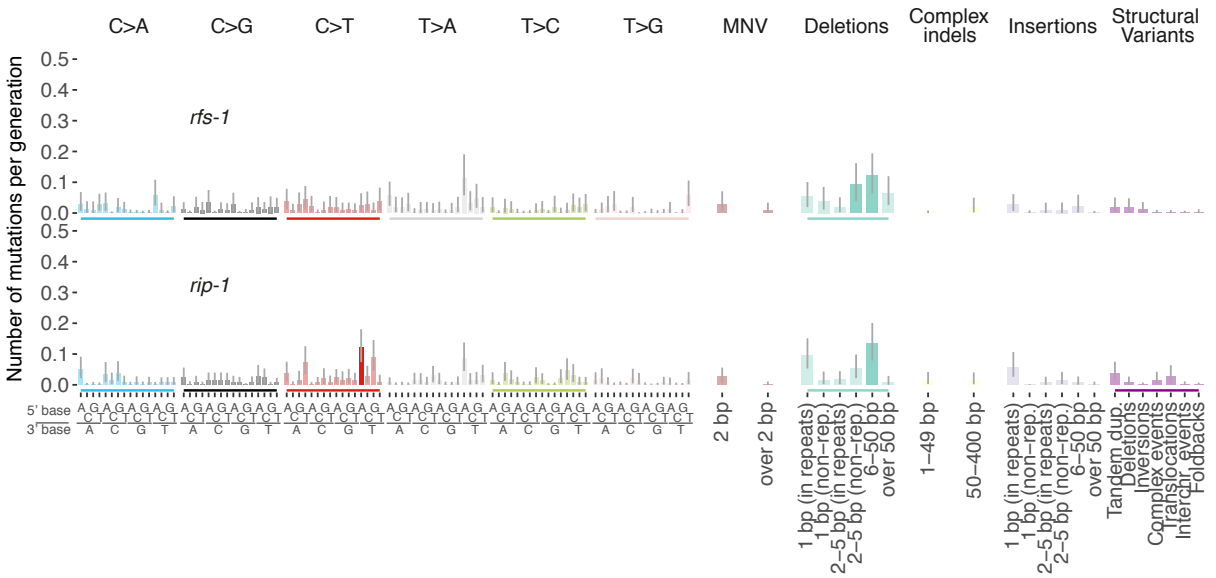


Figure 3.9: Experimental mutational signatures of the *rfs-1* and *rip-1* knockouts deficient in essential HRR nucleases. Brighter colors represent mutation contexts which are different from the wild-type. Lines underneath denote mutation classes where the total number of mutations within this class is different from the wild-type (FDR 10%).

DNA tails. The knockouts of the essential nucleases in *C. elegans*, *rad-51*, *mre-11*, and *com-1* are sterile due to defects in meiotic recombination. We thus analysed strains defective for the *rad-51* paralog *rfs-1*, and a knockout of *rip-1*, which encodes an RFS-1 interacting protein. The RFS-1/RIP-1 complex is required to remodel presynaptic RAD-51 containing filaments and to facilitate stand invasion (Taylor et al. 2015).

The strains with defective *rfs-1* and *rip-1* both showed a two-fold increase in mutagenesis (Figure 3.2). Mutational signatures of these knockouts both displayed an increased number of small deletions (Figure 3.9). Apart from small and not biologically relevant fluctuations (somewhat elevated level of SVs in *rip-1* deficient strains), the mutational spectra accumulated in the respective mutants were similar to each other and the signature of *brc-1* deficiency, consistent with RFS-1/RIP-1 being an essential HRR factor as well as BRC-1.

Cohesin complex and DNA stabilisation. Another complex involved in HR is a ring-shaped cohesin complex considered to tether broken DNA strands with the repair template on the sister chromatid to facilitate HR. Knockouts of *smc-5* and *smc-6*, which encode two components of this complex, both showed increased mutagenesis and a distinct mutational pattern characterised by elevated levels of SVs, predominantly deletions, tandem duplications and complex rearrangements (Figure 3.10). In line with the high preponderance of SVs in these strains, they could not be propagated beyond 5 generations when they became sterile. The signatures of these knockouts differ from *brc-1*

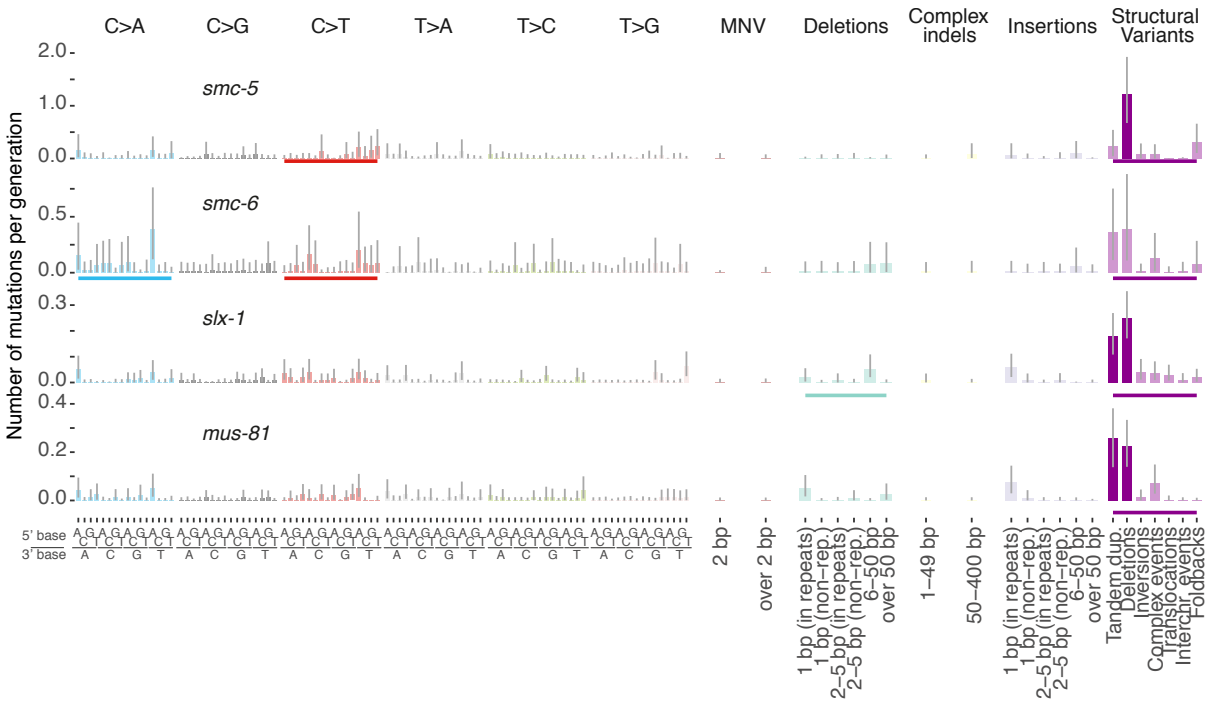


Figure 3.10: Experimental mutational signatures of the *smc-5*, *smc-6*, *slx-1* and *mus-81* knockouts, deficient in the cohesin complex or Holliday Junction-resolving nucleases. Brighter colors represent mutation contexts which are different from the wild-type. Lines underneath denote mutation classes where the total number of mutations within this class is different from the wild-type (FDR 10%).

signature through a high contribution of 10-kbps long deletions (Figure 3.7b). Given the SMC's role in stabilising the two sister chromatids, it is likely that the HRR is initiated, but the D-loop falls apart prematurely due to the instability of the whole structure, resulting in the loss of genetic material.

Holliday Junction resolution. Next, the invading strand has to capture the second end of the double-strand break upon the D-loop extension. After this, the two DNA double strands engaged in HR form cruciform four-way DNA intermediates referred to as Holliday Junctions, which need to be resolved by DNA structure-specific nucleases such as MUS81 and SLX1. *C. elegans slx-1* and *mus-81* mutants showed an identical mutational spectrum, characterised by an increase in structural variant rates, especially tandem duplications (similar to *brc-1*) and deletions with a median size of 7-8 kbps (Figure 3.7b, 3.7a, 3.10).

Helicases preventing wrong template choice. Another crucial factor of homologous recombination repair are helicases, enzymes that unwind double-stranded DNA. They unwind the D-loop structures when the sequence of the invading strand does not perfectly match the template strand, thus preventing recombination with homeologous template sequences. To investigate the patterns induced by helicase defects, we anal-

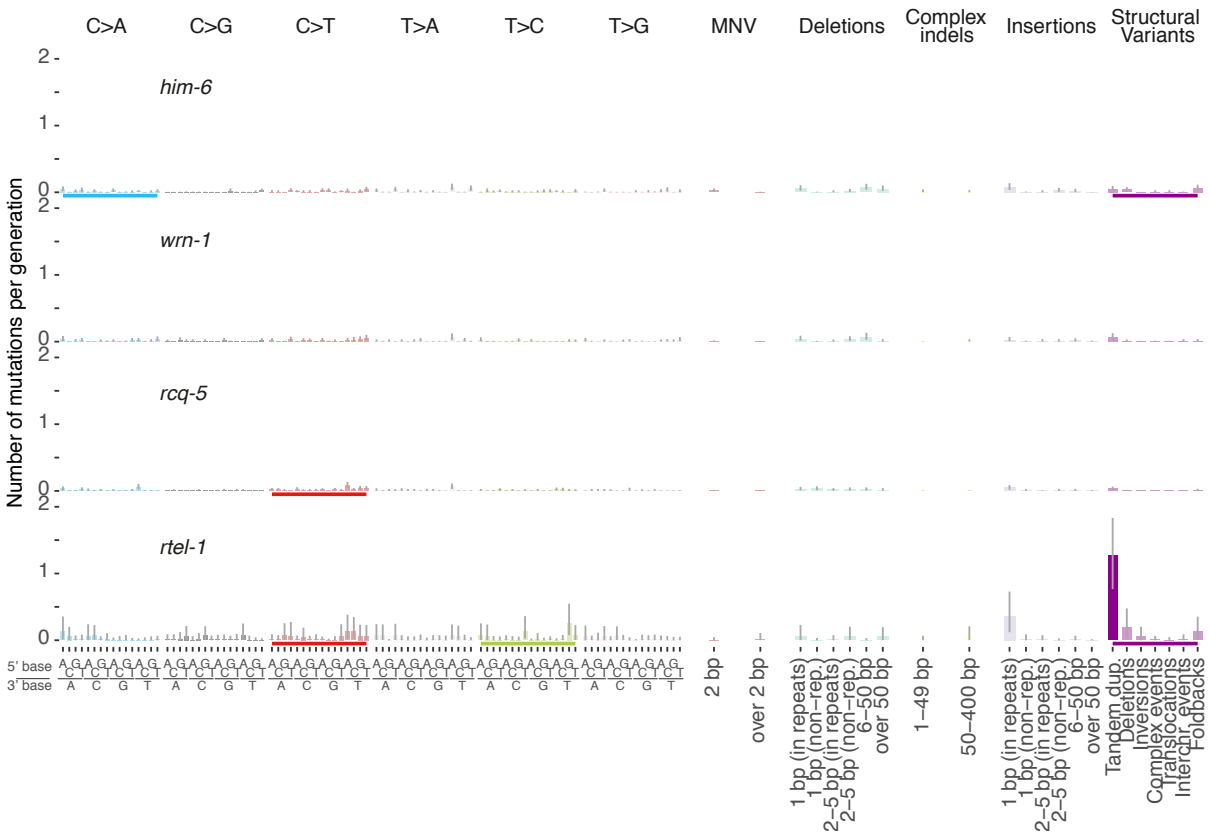


Figure 3.11: Experimental mutational signatures of the *him-6*, *wrn-1*, *rcq-5* and *rtel-1* helicase knockouts. Brighter colors represent mutation contexts which are different from the wild-type. Lines underneath denote mutation classes where the total number of mutations within this class is different from the wild-type (FDR 10%).

ysed the 3 *C. elegans* RecQ helicases: *him-6* - an ortholog of human Bloom syndrome gene, *wrn-1* - an ortholog of Werner's syndrome gene, *rcq-5* helicase, and also *rtel-1*, a conserved helicase involved in genome stability and telomere maintenance. Consistent with the similar function of these helicases, they showed spectra similar to each other. The only noteworthy change was a high rate of tandem duplications in *rtel-1* mutants, which made its mutational signature more similar to *brc-1* mutants (Figure 3.7a, 3.11). RTEL-1 helicase differs from other helicases in its ability to counteract recombination and promote synthesis-dependent strand annealing (Uringa et al. 2010); the difference in mutational spectra likely reflects the pathogenicity of extensive recombination for the repair of spontaneous DSBs.

Overall, our results recreate the mutational signatures of different kinds of homologous recombination deficiency. The distribution of mutations and the presence of microhomologies indicate NHEJ and MMEJ, as well as a failed HRR, as the sources of mutations observed upon propagation of HR deficient lines.

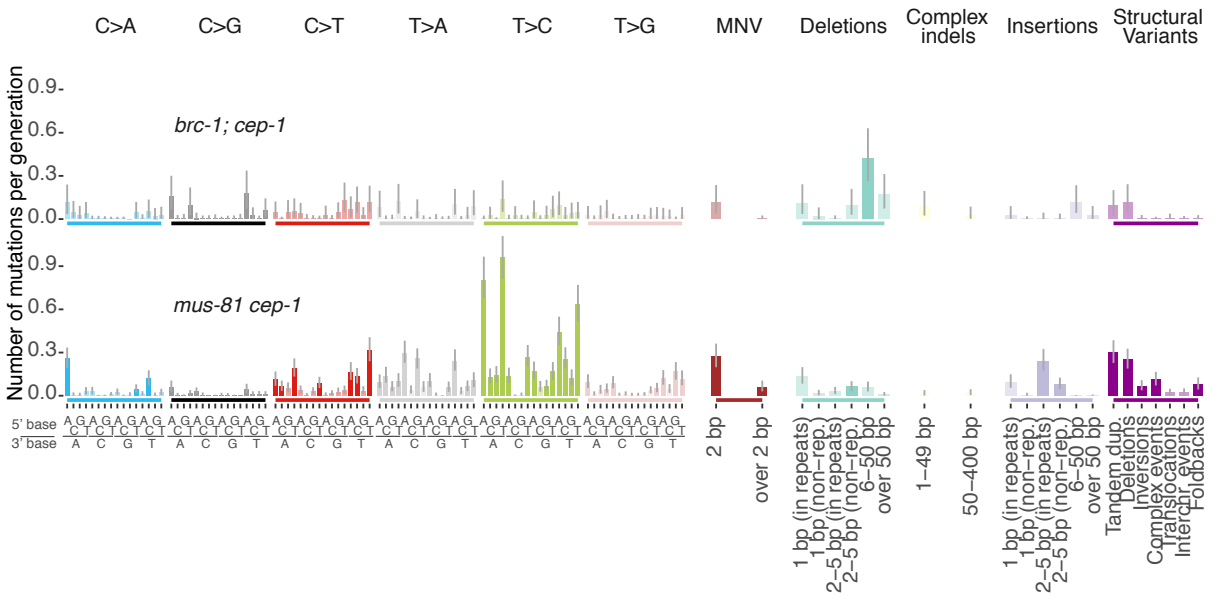


Figure 3.12: Experimental mutational signatures of the *brc-1; cep-1*, and *mus-81; cep-1* double knockouts. Brighter colors represent mutation contexts which are different from the wild-type. Lines underneath denote mutation classes where the total number of mutations within this class is different from the wild-type (FDR 10%).

3.3.5 Defects in DNA damage signalling exaggerate mutagenesis upon HR deficiency

A cell's reaction to DNA damage starts with damage signalling. Functioning DNA damage checkpoints are required to delay cell cycle progression and facilitate DNA repair. When the damage is excessive, DNA damage checkpoint can also trigger apoptosis to remove abnormal and defective cells. In fact, the key human apoptosis regulator, *TP53*, is mutated in more than half of cancers (Knijnenburg et al. 2018) which allows the cell to accumulate deleterious mutations inducing fitness and metabolic changes.

Analysing mutants defective for *atm-1*, encoding for *C. elegans* ATM, a PI3-kinase involved in detecting DSBs, we found a two-fold elevation of the level of structural variants, but no specific change in SV type preference. Knockouts of *cep-1*, *ced-3* and *ced-4* showed no effect at all (Figure 3.2). *cep-1* is the *C. elegans TP53* homolog and is required for DNA damage-induced apoptosis, while *ced-3* and *ced-4* encoding for a caspase and an Apaf-1 like protein, respectively, are required for both DNA damage induced as well as developmental apoptosis.

Observing distinctive patterns of mutations upon homologous recombination deficiency, we wanted to see if they may be exaggerated by knocking off DNA damage response components. Double knockouts of apoptosis genes and *him-6* helicase did not show any extreme effects beyond those induced by *him-6* alone (Figure 3.2, **??**). However, double

knockouts of *cep-1* gene with *brc-1* and with *mus-81* produced signatures different from the spectra accumulated in the single mutants (Figure 3.12).

Absence of CEP-1 exaggerated the mutational spectra induced by *brc-1* deficiency, leading to a 2-fold increase in the rate of base substitutions and small indels compared to a single HRR mutant (Figures 3.2). Furthermore, a double knockout of *mus-81* helicase and *cep-1* apoptosis regulator boosted the mutation rate about 10 times, creating a specific pattern of T>N and multinucleotide substitutions, small insertions, as well as large tandem duplications and deletions (Figure 3.12). Interestingly, the hypermutant profile reflected in the signature was present in only one of the two independently created strains, indicating that there may have been an additional factor triggering excessive mutagenesis in these samples. Moreover, the contributing mutations were concentrated in well-defined clusters (Figure 3.13).

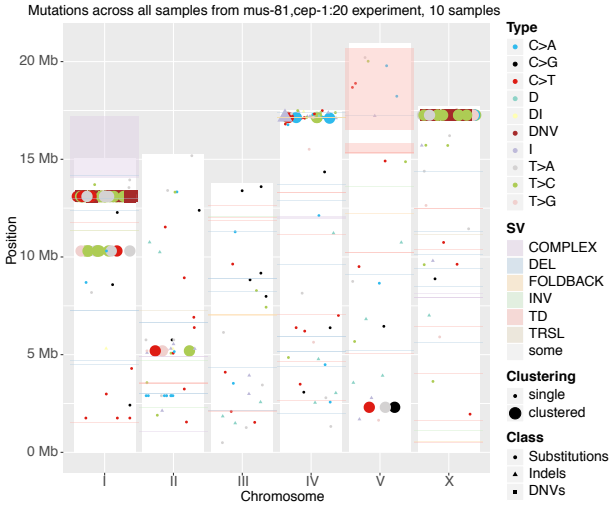


Figure 3.13: Distribution of mutations across 10 *mus-81; cep-1* double mutants.

3.3.6 Clustering of mutations across genotypes

Genomic clustering of mutations is a quite severe phenotype which would be selected against in fully DNA repair and damage signalling proficient cells unless they are under selective pressure for higher diversity such as in case of somatic hypermutation of immunoglobulins in immune cells (Chahwan et al. 2012). However, localised types of damage, such as double-strand breaks, can produce clusters of mutations upon error-prone repair. In addition, prolonged exposure of single-stranded DNA to the cell environment during replication or transcription can lead to clusters of mutations, which are created via repairing the damage incurred by antiviral protection mechanisms or secondary structure formation (Chan and Gordenin 2015).

We investigated the numbers of clusters and proportions of clustered mutations across genotypes and identified 7 backgrounds associated with an elevated rate of mutation clusters (Figure 3.14). In line with the expectation, most of these were associated with homologous recombination deficiency (*brc-1*, *him-6*, *rip-1* and *rfs-1*), which leads to DSBs being repaired by MMEJ or NHEJ pathways leaving clustered mutations close to the

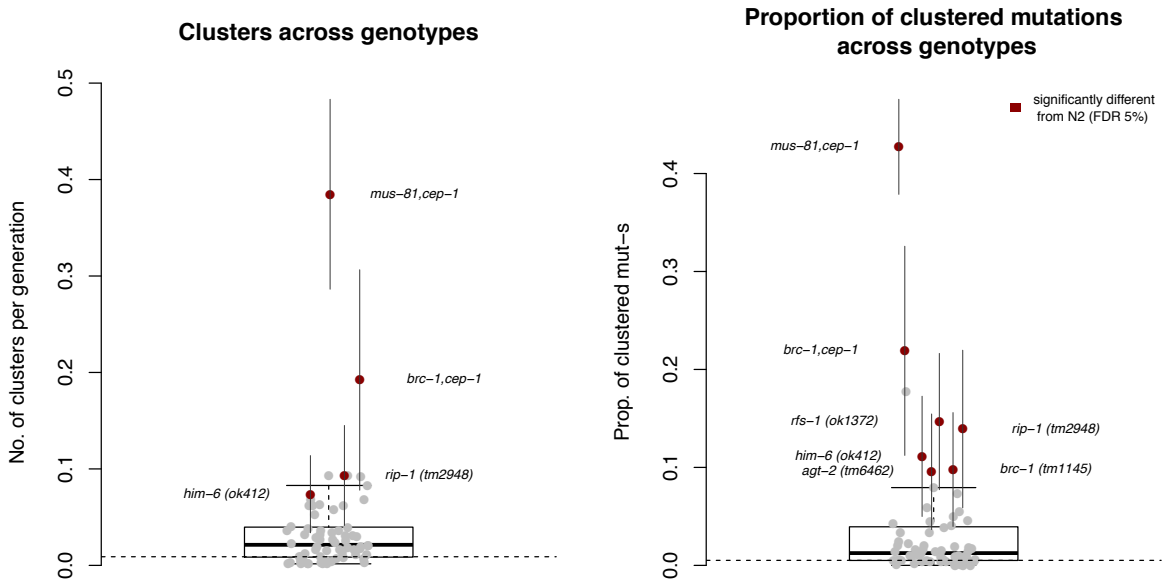


Figure 3.14: Numbers of clusters per generation for each genotype (left) and the fractions of clustered mutations per genotype (right). Dotted line marks the wild-type level.

breakpoints.

mus-81; cep-1 double knockouts demonstrated the highest amount of clusters and mutations within those clusters. As mentioned above, the clustering and hypermutation occurred simultaneously and only in half of the samples, which indicates that there may have been an additional mutational event leading to a leap in mutation rates and formation of dense mutational clusters.

Another HR and apoptosis deficient mutant line, *brc-1; cep-1*, had on average 20% of its mutations belonging to clusters which were arising a rate of 0.2 per sample per generation. Similar to the mutational spectra, an additional *cep-1* knockout seems to exaggerate the features of a single *brc-1* knockout, which also demonstrate cluster formation covering on average 10% of mutations. Thus, at least for HR deficiency conferred via *brc-1* knockout, the additional *cep-1* knockout allows for tolerating more mutations and more intensive clustering of these mutations. Benefits of such combinations for mutagenesis can be detected in cancers as well: breast cancers with germline *BRCA1/BRCA2* defects were shown to be enriched with somatic *TP53* mutations compared to sporadic tumours (Greenblatt et al. 2001).

An unexpected background in this list is the *agt-2* knockout. AGT-2 is one of the two predicted DNA alkyl-guanine alkyl-transferases in *C. elegans*, enzymes performing direct damage reversal upon alkylation of guanines. Mutational clusters were detected

in several independent samples and came together with a modest 2-fold increase in the base substitution rate (Figure 3.2). One of the possible explanations may be the ability of O6-alkylguanine to serve as a template for a TLS polymerase acting at a site of localised damage, leading to C>T changes on the opposite strand.

3.4 Discussion

In this chapter, I catalogued the mutational characteristics of DNA repair deficiencies across several DNA repair-related pathways in *C. elegans*. High-resolution characterisation of mutation rates, mutational signatures and local features provided insights into mechanisms of mutation acquisition. We described the process of deletion accumulation in GC-rich sequences for *dog-1* mutants, and replication-associated aggregation of tandem duplications in repetitive regions for *helq-1* helicase mutants. Moreover, our data suggested that localisation of mutational clusters in any other genomic region is a highly deleterious trait which would normally lead to cell death upon apoptosis induction.

The absence of significant mutagenic effects for the majority of genotypes indicates a high level of redundancy among different DNA repair pathways, which means that it may require the combined deficiency of multiple DNA repair pathways to trigger excessive mutagenesis. Equally, a latent defect in DNA replication integrity might only become apparent in conjunction with a DNA repair deficiency. Indeed, the increased mutation burden detected in the *pole-4*; *pms-2* double mutant while no increased mutation rate is observed in *pole-4* alone suggested a latent role of *pole-4*. Similarly, double knockouts of HR repair and apoptosis-related genes, *brc-1*; *cep-1* and *mus-81*; *cep-1* reveal the aggregation of clustered mutations upon alternative end-joining repair of DSBs which is only visible in cells lacking the apoptosis regulators.

However, when we do observe an effect different from that in the wild-type, we often do not know where these mutational spectra come from. It is clear that the damage which produces mutations upon an absence of mismatch repair is incurred by the polymerases which incorporate wrong or damaged bases; for other DNA repair pathways, not directly liked to replication, it is much harder to infer the original damaging agent. There are many processes which can cause chemical or mechanical damage to the cells without any exposure to exogenous mutagens: segregation, cell movement, replication-transcription collisions, meiosis, mechanical stress.

In the next chapter, I will study the mutational signatures of mismatch repair deficiency in more detail, and show the usability of *C. elegans* research to understanding the aetiology of mutations in human cancers.

Chapter 4

Comparison of mutational signatures of mismatch repair deficiency in *C. elegans* and human gastrointestinal cancers

4.1 Introduction

In the previous chapter, I showed that many types of DNA repair deficiency could alter spontaneous mutation rates and generate specific mutational patterns. In particular, the mutants with defective mismatch repair have drawn our attention as they had the most well-defined mutational spectra and the highest amount of mutations across the genotypes.

One of the initial questions that this study aimed to answer was whether the mutational signatures are consistent across species, and how well the signatures obtained from model systems could explain the mutations observed in cancer. As MMR deficiency is also a well-known carcinogenesis factor (Hsieh and Yamane 2008), the MMR mutants were the best candidates to explore the translational potential of *C. elegans* derived signatures.

Moreover, mismatch repair deficiency has been associated with more than one mutational signatures (Alexandrov et al. 2013b, Alexandrov et al. 2018). Adding the information from a model organism, where mutagenesis was performed in a completely controlled environment, may help to disentangle the relationship between these different mutational spectra and the underlying mutational processes.

In this chapter, I will present a comparison of MMR deficiency between *C. elegans* and human cancer data, and study the composition and origins of mutations observed in MMR deficient cancers.

Contributions

The following work was published in co-authorship with Bettina Meier and colleagues:

Meier, B., Volkova, N.V., Hong, Y., Schofield, P., Campbell, P.J., Gerstung, M., and Gartner, A. 2018. Mutational signatures of DNA mismatch repair deficiency in C. elegans and human cancers. *Genome Research* **28**, 666-675.

For this project, AG, MG and PJC have conceived the study, BM and PS prepared the *C. elegans* data, and BM and YH and suggested initial filtering, mutation rate analysis and homopolymer analysis. I extracted the mutational signatures from experimental data, derived the final estimates for indels per homopolymer, and preformed the comparative study in human cancers (including signature analysis, association with indels and search for interaction factors). Compared to the publication, this chapter focuses on the comparison between *C. elegans* and cancer, and includes an additional adjustment of single-base indels when comparing mutational signatures of MMR deficiency between species. All the scientific results, as well as figures and figure captions, were published in Meier et al. 2018.

4.2 Mismatch repair deficiency in cancer

Mismatch repair (MMR) was one of the first DNA repair pathways to be associated with cancer predisposition: mutations in MMR genes were associated with hereditary non-polyposis colorectal cancer (HNPCC), also referred to as Lynch Syndrome, at the beginning of 20th century (Fishel et al. 1994, Bronner et al. 1994, Nicolaides et al. 1994, Papadopoulos et al. 1994, Miyaki et al. 1997). Defects in MMR related genes were found to be the cause of more severe conditions leading to high rates of gastrointestinal, endometrial and brain cancers: biallelic mismatch repair deficiency syndrome (bMMRD) (Durno et al. 2015), Muir-Torre syndrome and Turcot's syndrome (Lawes, SenGupta, and Boulos 2003).

Several cancer types have a high frequency of MMR deficiency (Figure 4.1), especially so uterine, stomach and colorectal cancers (Cortes-Ciriano et al. 2017). Among sporadic colorectal and gastric cancers, about 15% harbour MMR defects (Funkhouser et al. 2012, Cancer Genome Atlas Research Network 2014a). This number is higher in uterine cancers (up to 33% (Hecht and Mutter 2006)) but much lower in other cancer types (Cortes-Ciriano et al. 2017, Bonneville et al. 2017).

MMR deficiency in humans (as well as many model organisms) mainly manifests via contraction and expansion of repetitive regions called microsatellite instability or short tandem repeats, which would normally be repaired by MMR machinery (Denver et al.

2004, Hanford et al. 1998, Lujan, Clark, and Kunkel 2015). As discussed in the Introduction, three mechanisms ensure the fidelity of replication. Selectivity towards the correct nucleotide and 3'-5' exonuclease ability of the main replicative polymerases prevents some of the errors. However, the misincorporation rate is still at the order of 10^{-5} potentially leading to thousands of mismatches genome-wide turning into mutations during the next round of replication (Kunkel and Bebenek 2000).

In healthy human cells, postreplicative mismatch repair repairs these mismatches on the newly synthesised strands, reducing the resulting mutation rate down to about 10^{-10} corresponding to 1 mutation per cell division (Bernstein et al. 2013). Upon a knockout of an MMR gene, the mutation rate can increase up to 100-1000 fold (Simpson 1997, Zou et al. 2018). The fold-change in cancer is more moderate as cancers generally have a higher mutation rate: MMR proficient colorectal adenocarcinomas have an average point mutation rate of ∼4 mutations per Mb, whereas in MMR-deficient colorectal cancers it is 10 times higher and reaches 40-50 mutations per Mb (Cancer Genome Atlas Research Network 2012).

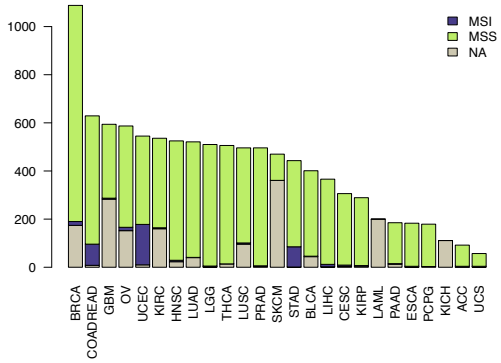


Figure 4.1: Number of tumours per cancer type in TCGA with experimentally identified or predicted microsatellite instability (MSI). MSS - microsatellite stable, NA - data not available. Based on the data from Lee et al. 2015 and Cortes-Ciriano et al. 2017.

Before the spread of high-throughput sequencing, the detection of mismatch repair deficiency was performed based on measuring the number of repeats in five selected microsatellites and comparing them between the tumour and normal samples (Boland et al. 1998, Laghi, Bianchi, and Malesci 2008). Cancer predisposition used to be identified by detecting specific single-nucleotide polymorphisms in MMR genes (Buhard et al. 2006). Following the aggregation of sequencing data, a number of computational methods were developed that classify the samples based on different data types (whole genome, whole exome, or targeted sequencing) and characteristics of cancer samples: mutational burden, number of indels in repetitive regions, and density of mutations in particular microsatellites (Cortes-Ciriano et al. 2017, Niu et al. 2014, Bonneville et al. 2017, Huang et al. 2015, Nowak et al. 2017). Another typical marker used in the clinic is the epigenetic silencing of *MLH1*, which was found to be common in cancers with sporadic MMR deficiency and is easier to reliably detect via immunohistochemistry (Herman et al. 1998, Boissière-Michot et al. 2016).

Identification of functional MMR deficiency in tumours is important for making deci-

sions about treatment and prognosis. It has been associated with better clinical outcome (Lee et al. 2002) and better response to immunotherapy (Dudley et al. 2016, Le et al. 2015, Kelderman, Schumacher, and Kvistborg 2015).

4.3 Mutational spectra of mismatch repair deficiency in *C. elegans*

Compared to human, the *C. elegans* genome does not encode obvious MutL-β and γ sub-units (PMS1 and MLH3 homologs, respectively), while the homologs for MutL-α sub-units *MLH1* and *PMS2* can be readily identified using homology searches. We studied two knockouts lines, *mlh-1* and *pms-2* mutants.

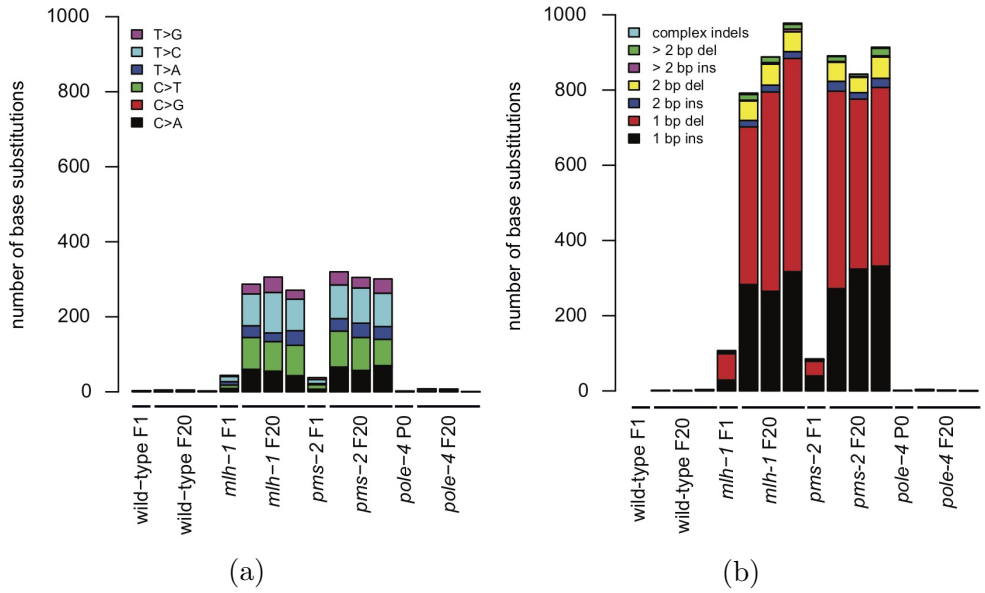


Figure 4.2: The numbers of single-base substitutions (a) and indels (b) in the samples with *mlh-1* or *pms-2* knockouts of different generations.

4.3.1 Mutation types and rates in MMR mutants

The mutational spectra of the two knockouts were primarily defined by C>A, C>T, T>C substitutions as well as single-base deletions and insertions (Figure 4.2, 4.3). A similar prevalence of T>C and C>T transitions was previously reported in *S. cerevisiae msh2* mutants and in MMR defective human cancer lines (Alexandrov et al. 2013b, Lujan et al. 2014, Supek and Lehner 2015). Analysing these base substitutions within their 5' and 3' sequence context, we found no enrichment of distinct 5' and 3' bases associated with T>C transitions prominent in *mlh-1* and *pms-2* single mutants. In contrast, T>A

transversions occurred with increased frequency in an ATT context, C>T transitions in a GCN context, and C>A transversions in an NCT context (Figure 4.3).

Analysis of the broader sequence context of T>A transversions in an ATT context revealed that > 90% of substitutions occurred in homopolymer sequences; the majority (> 75%) in the context of two adjoining A and T homopolymers. Similarly, an increased frequency of base substitution at the junction of adjacent repeats has recently been reported in *S. cerevisiae* MMR mutants, giving rise to the speculation that such base substitutions may be generated by double slippage events (Lang, Parsons, and Gammie 2013). To further analyse base changes, we visually searched for base changes occurring in repeat sequences. We found several examples in which one or several base substitutions had occurred that converted a repeat sequence such that it became identical to flanking repeats consistent with polymerase slippage across an entire repeat. Such mechanisms could lead to the equalization of microsatellite repeats - a phenomenon referred to as microsatellite purification (Harr, Zangerl, and Schlötterer 2000).

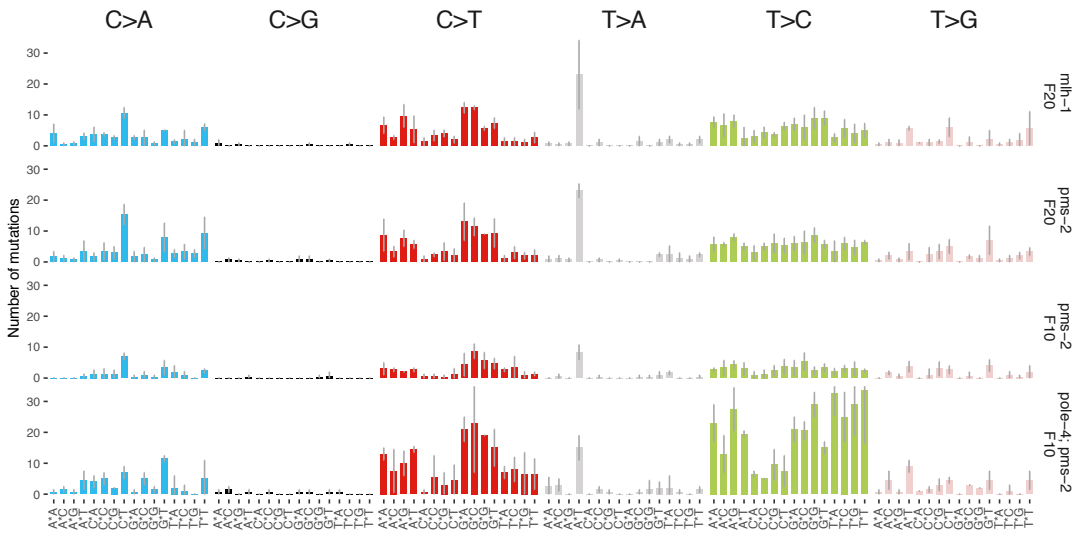
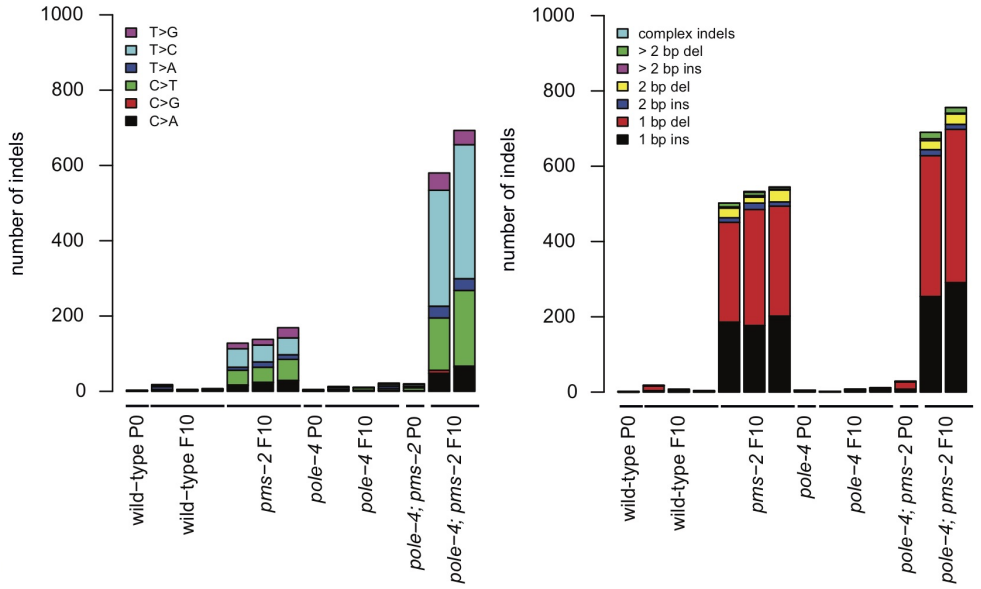


Figure 4.3: Spectra of base substitutions in *mlh-1*, *pms-2* and *pole-4; pms-2* knockout lines of different generations, along with their 95% confidence interval. Corresponds to Figure 1A,B in Meier et al. 2018.

The genome-wide mutation rates observed in the absence of *C. elegans* MutL-α proteins MLH-1 and PMS-2 agree with mutation rates previously determined for *C. elegans* MutS and *S. cerevisiae* MMR mutants (Strand et al. 1993, Yang et al. 1999, Degtyareva et al. 2002, Tijsterman, Pothof, and Plasterk 2002, Denver et al. 2005). The mutational signatures extracted from the two knockouts were remarkably similar (cosine similarity of 0.97, Figure 3.3). However, similar experiments in mammalian cells (Yao et al. 1999, Baross-Francis et al. 2001) demonstrated difference in mutation rates and spectra for cells deficient in *MLH1* or *PMS2*, which suggests that the inactivation of the MutL-α

heterodimer in *C. elegans* is sufficient to yield a fully penetrant MMR deficient phenotype. This observations is consistent with the absence of any *PMS1* MutL-β and *MLH3* MutL-γ homologs in the *C. elegans* genome.

4.3.2 Interaction between MMR and pol ε



(a) (b)

Figure 4.4: The numbers of single-base substitutions (a) and indels (b) in the samples with *pms-2* single or *pole-4; pms-2* double knockouts of different generations. Corresponds to Figure 1C,D in Meier et al. 2018.

Mismatch repair ought to repair the errors left by a replicative polymerase. Deficiencies in human polymerase epsilon and polymerase delta were previously reported as driver events in hypermutated brain cancers, as well as colorectal and endometrial cancers (Shlien et al. 2015, Shinbrot et al. 2014). Given that the null alleles of the human and *C. elegans* leading strand polymerase Pol ε catalytic subunit, *POLE* and *pole-1*, respectively, are essential for viability, we focused our analysis on a non-essential *C. elegans* Pol ε subunit, termed POLE-4. Dbp3p, the *S. cerevisiae* POLE-4 ortholog, was shown to be implicated in the stabilization of POLE with the primer-template DNA complex (Aksenova et al. 2010).

pole-4; pms-2 mutants could not be readily propagated beyond the F10 generation, hence we compared them to wild-type, *pms-2*, and *pole-4* mutant strains which were also grown for 10 generations. *pms-2* mutant strains carried an average of 145 base substitution and 527 indels over 10 generations, roughly half the number we observed in the F20 generation (Figure 4.4). In comparison, the number of single-base substitutions and

indels was increased 4.4 fold and 1.4 fold in *pole-4; pms-2* double mutants, respectively (Figure 4.4).

In contrast to *mlh-1* and *pms-2* mutants, *pole-4* single knockout lines exhibited mutation rates and profiles not significantly different from the wild-type - a finding which was confirmed by propagating *pole-4* propagated over 40 generations (Figure 3.2). However, the double *pole-4; pms-2* mutants demonstrated a 2-fold increase in total mutation rate over the single MMR mutants.

The fact that *pole-4; pms-2* mutants could not be grown beyond F10 suggests that a mutation burden higher than 500-700 single-base substitutions in conjunction with the 700-800 indels might be incompatible with organismal reproduction. These numbers are in line with our inability to propagate *mlh-1* and *pms-2* single mutant lines for 40 generations. The multiplicative effect on mutation burden detected in *pole-4; pms-2* double mutants alongside with unchanged mutation rate for *pole-4* alone suggests that replication errors occur at increased frequency in the absence of *C. elegans pole-4* but are effectively repaired by MMR.

Our finding that *pole-4* mutants do not show increased mutation rates is surprising given that the deletion of the budding yeast POLE-4 homolog Dpb3 leads to increased mutation rates comparable to the proof-reading deficient pol2-4 allele of the Pol ε catalytic subunit (Aksenova et al. 2010, Lujan et al. 2012). Increased mutation rates have also been reported for proof-reading defective POLE Pol ε catalytic subunit in mice and human cells, and in humans, such mutations are associated with an increased predisposition to colorectal cancer (Albertson et al. 2009, Palles et al. 2013).

While we could not define mutational patterns specifically associated with *pole-4* loss due to the low number of mutations, the profile of *pole-4; pms-2* double mutants differed from mismatch repair single mutants. Most strikingly, in addition to C>T transitions in a GCN context, T>C transitions were generated with higher frequency accounting for > 50% of all base changes (Figure 4.3). Interestingly, T>C substitutions were underrepresented in the context of a flanking 5' cytosine. Notably, T>C changes not embedded in a clearly defined sequence context have also been reported for MMR-deficient tumour samples containing mutations in the lagging strand polymerase Pol δ (Shlien et al. 2015), but not in *S. cerevisiae* and human tumours with a combined MMR and Pol ε deficiency (Lujan et al. 2014, Shlien et al. 2015)

4.3.3 Indels in homopolymeric sequences

The mutational spectra in *mlh-1* and *pms-2* single and *pole-4; pms-2* double mutants were mostly composed of small insertions/deletions (indels) (Figures 4.2, 4.4). Mismatch

repair is known to manifest via microsatellite instability expressed as expansion or contraction of highly repetitive regions, which led is to investigate the local context of indels in these mutants.

To assess the likelihood of these mutations falling into repetitive regions by chance, we calculated the total numbers of homopolymers, di- and tri-nucleotide runs encoded in the *C. elegans* genome, defined here as repetitive DNA regions with a consecutive number of identical bases or repeated sequence of $n \geq 4$ were identified from the reference genome WBcel235.74 using repeat search. As a result, we identified 3.433.785 homopolymers, with the longest homopolymer being comprised of 35 Ts in the *C. elegans* genome (Figure 4.5). In addition, we found 25,156 dinucleotide repeats and 7,615 trinucleotide repeats. In total, homopolymers covered about 16% of the *C. elegans* genome, similar to humans (16.3% as calculated on hg19). The fraction of genome covered by repetitive sequences composed of 2-6 bp repeats was overall less than 1%; hence, we focussed our analysis on homopolymeric regions.

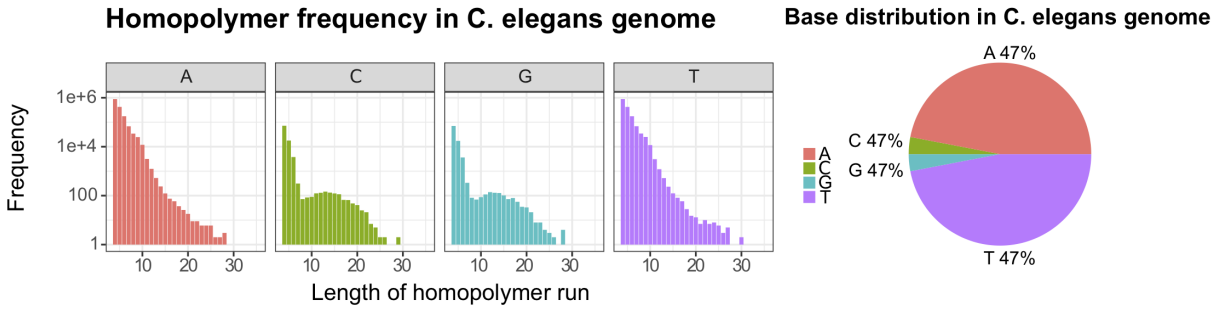


Figure 4.5: Distribution of homopolymer repeats encoded in the *C. elegans* genome by length and DNA base shown in log10 scale (left panel) and the relative percentage of A, C, G and T homopolymers in the genome (right panel). Corresponds to Figure 2A in Meier et al. 2018.

Overlaying the indels with a map of homopolymers showed that the absolute majority of indels in single and double mutant backgrounds, 90% of which were 1-bp insertions or deletions, occurred in homopolymer runs (Figures 3.3, 4.6). 2 bp indels accounted on

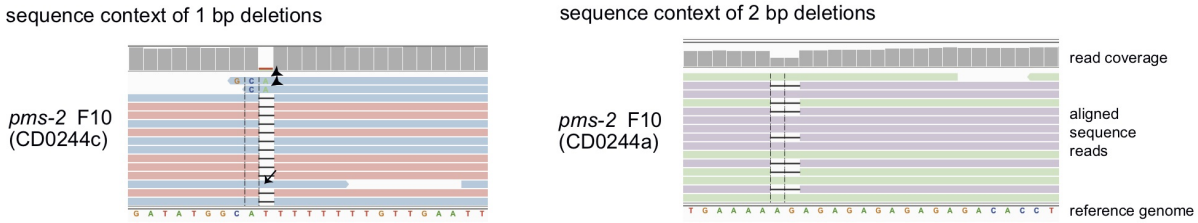


Figure 4.6: An example of sequence context of a single- and double-base indels in MMR deficient *C. elegans* mutants. Corresponds to Figure 1F in Meier et al. 2018.

average for 5.5-8.6% of all indels (Figure 4.2), and affected homopolymer runs as well as dinucleotide repeat sequences at a similar frequency. Similar results were also reported for MMR defective *S. cerevisiae* strains, human iPS cells and human organoids (Lujan, Clark, and Kunkel 2015, Zou et al. 2018, Drost et al. 2017).

Trinucleotide repeat instability has been associated with a number of neurodegenerative disorders, such as fragile X syndrome, Huntington's disease and Spinocerebellar Ataxias (Brouwer, Willemsen, and Oostra 2009). Based on our analysis, trinucleotide repeat sequences are present in the *C. elegans* genome at > 400 fold lower frequency than homopolymer runs. Across F20 and F10 generation samples of *mlh-1* and *pms-2* mutants, we observed between 3 to 7 trinucleotide indels per 10 generations predominantly in homopolymer sequences precluding estimation of mutation rates for these lesions.

Given the high number of indels arising in homopolymer repeats, we aimed to investigate the correlation between the frequency of indels and the length of the homopolymer in which they occurred. 47% of genomic homopolymers in *C. elegans* were comprised of As, 47% of Ts and 3% each accounted for Gs and Cs. A and T homopolymer frequencies decreased continuously with increasing homopolymer length; C and G homopolymer frequencies decreased up to homopolymer lengths of 8 bp, followed by roughly consistent numbers for homopolymers of 8-17 bp length and decreasing frequencies with longer homopolymers (Figure 4.5).

Although we identified slightly higher overall numbers of homopolymer runs in the genome, this base specific size distribution is consistent with previous reports (Denver et al. 2004). An average of ~ 0.5 1-bp indels arising in homopolymer sequences was observed in 101 wild-type lines of different generations, indicating the frequency with which such events might occur in wild-type or as amplification artefacts during sequencing.

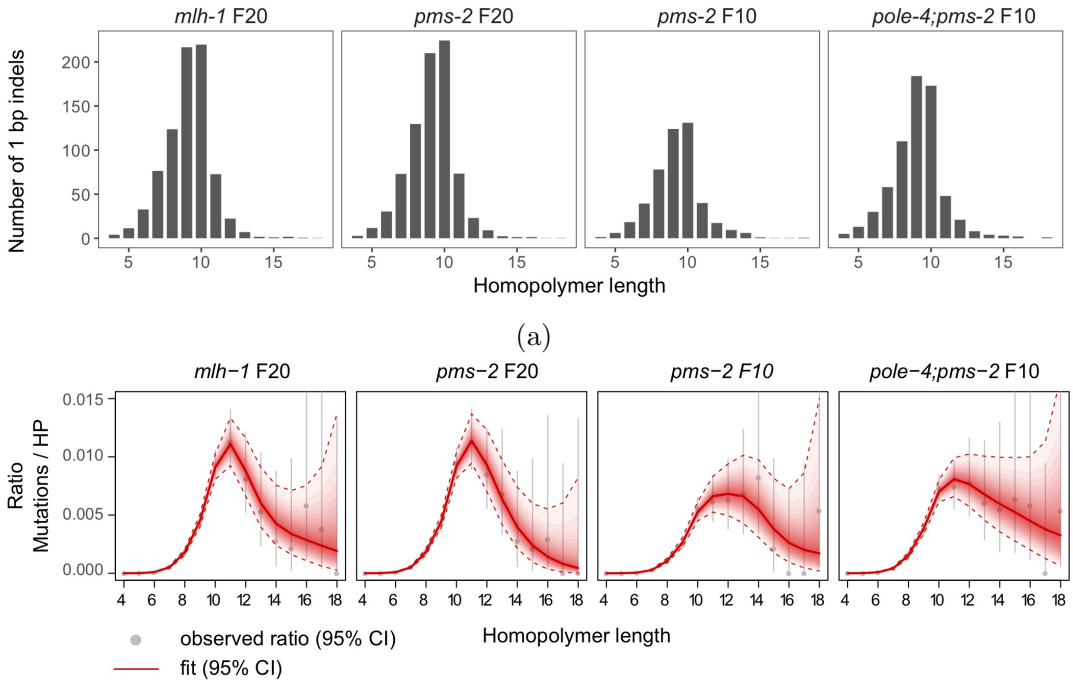
Small indels in repetitive sequences could be generated as polymerase errors during bridge amplification or sequencing, leading to possible sequencing artefacts. Across a total of 101 wild-type samples of different generations, 7,433 1-bp indels were observed on average prior to post-processing. Of these, 7,109 1-bp indels on average occurred in homopolymer runs. Following filtering, we observed on average 0.5 out of 0.54 1-bp indels arising in homopolymers per sample, with the majority of indels being removed when filtering for quality and frequency of mutant reads. Thus, 1-bp indels likely arising during PCR amplification or sequencing seem to be efficiently removed using our filtering procedure.

Plotting the frequency of all 1-bp indels observed in MMR deficient backgrounds in relation to the length of the homopolymer in which they occur, we found that the likelihood of indels increased with homopolymer length of up to 9-10 base pairs, and trailed off in longer homopolymers. Given that the frequency of homopolymer tracts decreases

with length, we normalised for homopolymer number (Figure 4.7a). These results are consistent with observations in budding yeast (Lang, Parsons, and Gammie 2013).

To assess the variability of the frequency estimation, we applied a generalised additive model (GAM) with a spline term which supported a rapid increase for homopolymers up to length 9 followed by a drop or plateau in indel frequency for longer homopolymer with decreasing confidence (Figure 4.7b). The lack of statistical power precludes firm conclusions about indel frequencies in homopolymers > 13 bp based on low numbers of long homopolymers in the genome and too few observed indel events (Figure 4.7).

In summary, our data suggest that replicative polymerase slippage occurs more frequently with increasing homopolymer length, with a peak for homopolymers of 10-11 nucleotides, followed by reduced slippage frequency in slightly longer homopolymers. A similar frequency distribution has been reported for human MLH-1KO organoids (Drost et al. 2017).



(a)

(b)

Figure 4.7: (a) Average number of 1-bp indels in homopolymer runs for different mutant lines adjusted for homopolymer length. (b) GAM fit for the ratio of 1bp indels normalised to the frequency of homopolymers (HPs) in the genome, best fit denoted by red line along with its 95% CI. The grey dots and grey bars indicate the average frequency and the 95% confidence intervals, respectively. Corresponds to Figures 2B-C in Meier et al. 2018.

4.4 Mutational processes shaping gastrointestinal tumours

4.4.1 De novo extraction of mutational signatures

The 2013 COSMIC v.2 mutational signature catalogue (Alexandrov et al. 2015, Forbes et al. 2015) contained 5 signatures associated with mismatch repair: 6, 15, 20, 21 and 26. It might indicate that mismatch repair deficiency contributes to mutagenesis via multiple mechanisms at the same time, but these signatures seemed to be uncorrelated across cancer types (Alexandrov et al. 2015). In order to gain insights into the aetiology of these signatures, we studied the mutational signatures of stomach and colorectal adenocarcinomas. Apart from having a high frequency of MMR deficiency, these two cancer types stem from similar tissues yet carry different mutational signatures associated with MMR deficiency (Alexandrov et al. 2013b).

The datasets acquired from ICGC (`http://icgc.org`) contained single nucleotide (SNV) and small indel variant calls from 215 and 289 donors, respectively. According to the TCGA Clinical Explorer (Lee et al. 2015), 40/215 and 63/289 samples were labelled as microsatellite instable-high (MSI-H), which we considered as an indicator of MMR deficiency. The samples labelled as microsatellite instable-low (MSI-L) or microsatellite stable (MSS) were considered MMR proficient. Comparison of the mutational spectra averaged across all MMR deficient samples within COAD and STAD datasets showed high similarity between the two cancer types, in contravention with the presence of different signatures in these cancers as determined by COSMIC analysis. The cosine similarity between them reached 0.98, with the main source of variability being the contribution of CpG>TpG (Figure 4.8). Based on this results, we hypothesised that the underlying signature of MMR deficiency in these two cancer types should be the same, but there may be different interacting processes present in these datasets that confine the mutational signature analysis.

In order to analyse the signatures of mismatch repair deficiency active in these cancer types, we performed an unsupervised signature extraction via non-negative matrix factorisation similar to that initially proposed by Alexandrov in Alexandrov et al. 2013b. However, the initial procedure was performed using Frobenius norm as the loss functions minimised during the optimisation, which did not take into account the discrete nature of mutation count data. Hence, we applied Brunet NMF with Kullback-Leibler divergence as an objective Brunet et al. 2004, which is equivalent to an additive Poisson model. Given the high prevalence of indels observed in *C. elegans* data, we also included single-base insertions and deletions to the count matrix used for signature extraction.

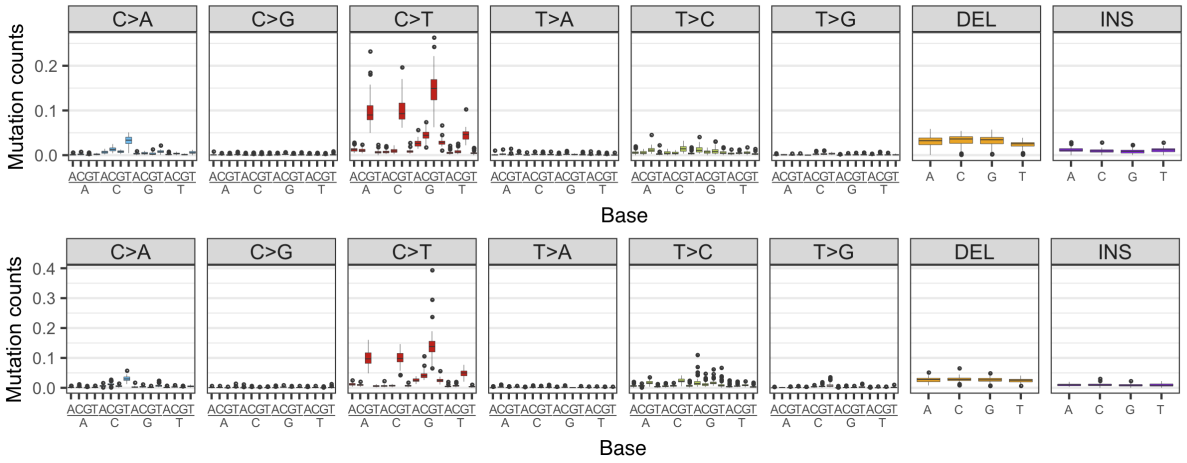


Figure 4.8: Averaged per-substitution type per-context variability of mutation spectra from MSI samples in COAD (top) and STAD (bottom) datasets. The averaged spectra look extremely similar (0.99 similarity between the per-type per-context means) with most of the variability coming from different fractions of C>T transitions in NCG contexts. Corresponds to Supplementary Figure 5C in Meier et al. 2018.

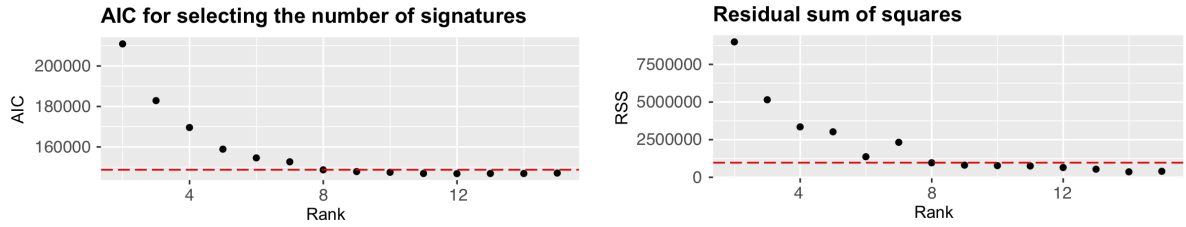


Figure 4.9: AIC and RSS for detecting the number of signatures in COAD/STAD dataset. Corresponds to Supplementary Figure 3A in Meier et al. 2018.

The number of signatures was chosen based on the saturation of both the Akaike Information Criterion (AIC) (Akaike 1992) and the residual sum of squares (RSS). The AIC is calculated as

$$AIC = 2k \cdot (n + N) - 2 \log L,$$

where the first term reflects the number of parameters (k is the number of dimensions, n is the number of signatures, N is the number of samples), and L denotes the maximised model likelihood. AIC penalises the As the L would naturally increase with the addition of parameters, we performed signature extraction for different ranks and chose the one where AIC and also RSS decrease slows down to avoid oversegmentation (Figure 4.9).

4.4.2 Aetiology of extracted signatures

Based on this metric, we extracted 8 mutational signatures from the combined STAD and COAD dataset (Figure 4.10). Many of these signatures matched to one or more

COSMIC signatures, validating these results.

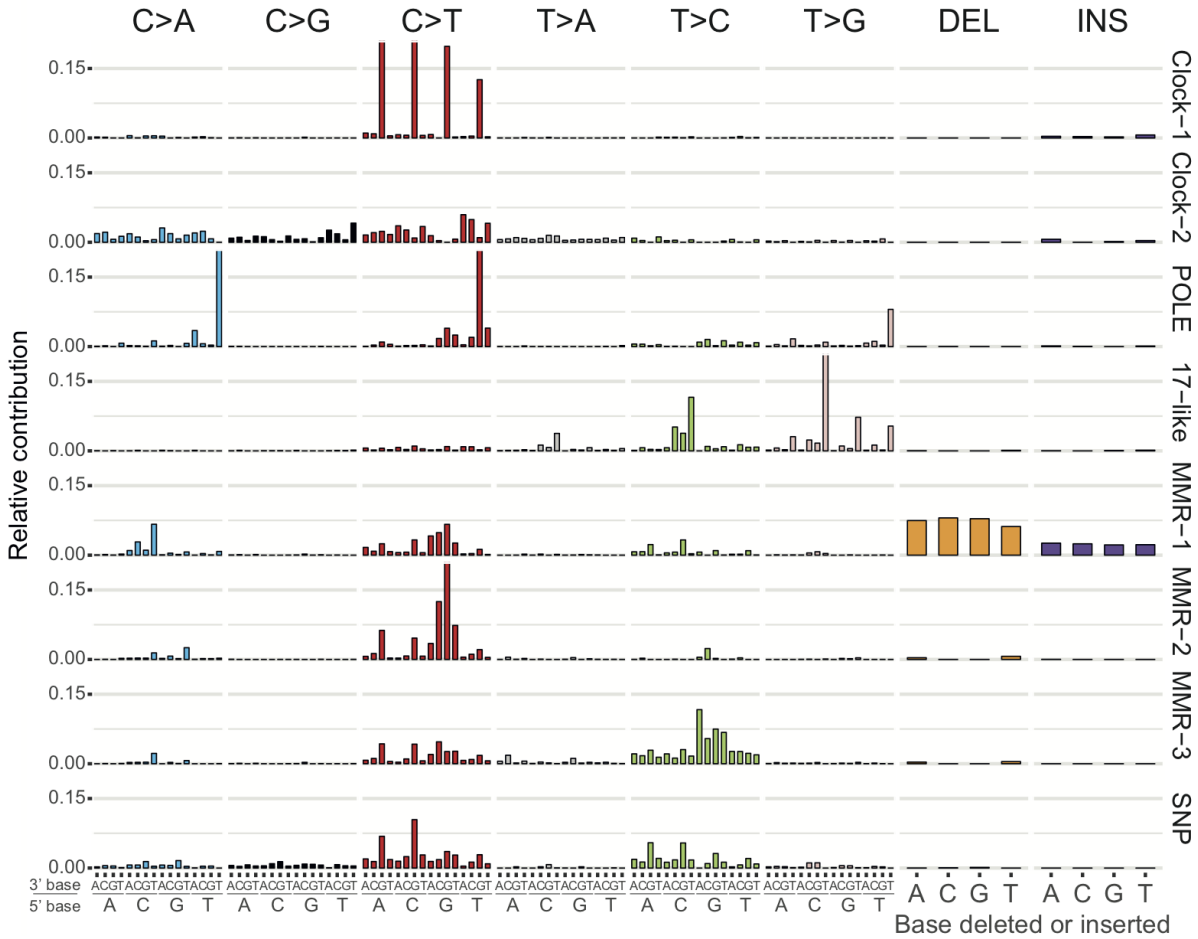


Figure 4.10: Mutational signatures including base substitutions and 1-bp indels derived from the combined COAD-US and STAD-US data sets. Corresponds to Figure 3B in Meier et al. 2018.

For three signatures out of eight, the fraction of mutations assigned to these signatures was significantly higher in MMR-deficient cancers compared to proficient ones: for MMR-1 (one-tailed t-test p-value $= 4.7 \cdot 10^{-55}$), MMR-2 (p-value $= 1.1 \cdot 10^{-11}$) and MMR-3 (p-value $= 6.0 \cdot 10^{-12}$) (Figure 4.11). Of these, MMR-1 mostly resembled signature 20, MMR-2 - signature 15, and MMR-3 - signatures 21 and 26. Interestingly, only one signature, MMR-1, carried all the indel information, which indicates that the profile of base substitutions defined by this signatures is likely to be the typical distribution of mutations stemming from replicative errors. This signature also showed high accuracy in classification of MMR deficiency (AUC - the probability of a random MSI sample having higher MMR-1 contribution than a random MSS sample - of 0.985).

Additional signatures identified in the tumour samples were characteristic of defects in the proofreading domain of replicative polymerase ε ("POLE") (Alexandrov et al.

2013b, Shinbrot et al. 2014). A signature characterised by C>T mutations in a CpG base context, which is most likely to result from 5-meC deamination (Alexandrov et al. 2015), was referred to as "Clock-1 (5meC)". Signature "Clock 2" was present in the majority of samples and likely reflected the background mutation rates (similar to COSMIC signature 5). "17-like" is a signature predominantly found in stomach cancers and highly similar to COSMIC signature 17 (Alexandrov et al. 2013b), which was suggested to be associated with some mutagenic exposure based on its transcriptional strand asymmetry (Tomkova et al. 2018), and was recently found to closely resemble mutational spectrum of 5-fluoracil treatment that depletes the pool of thymine triphosphates (Christensen et al. 2019).

Finally, we also identified a signature predominantly consisting of T>A mutations at CpG sites. This type of mutations was present in high fraction in a limited number of samples (all of which were coming from the same sequencing centre), which indicated an artefactual nature of this signature. Indeed, comparing the base substitutions from the samples carrying over 50% of mutations assigned this signature to the set of SNPs prevalent in the human population (based on the dbSNP database Sherry et al. 2001) showed a high overlap: the mutational burden of these samples consisted, on average, of 76% (IQR 76-85%) SNPs, whereas for the rest of samples this number was much lower (18%, with IQR 0-40%). In addition to that, we compared the ratio of coding and non-coding SNPs (The 1000 Genomes Project Consortium 2015) and identified a much lower ratio of non-synonymous to synonymous changes (0.8 with IQR of 0.68-0.88) across these samples than expected for cancer variants (higher than 1, Ding et al. 2008), suggesting a germline nature of these variants. Hence, this signature was referred to as "SNP" and considered as an artefact.

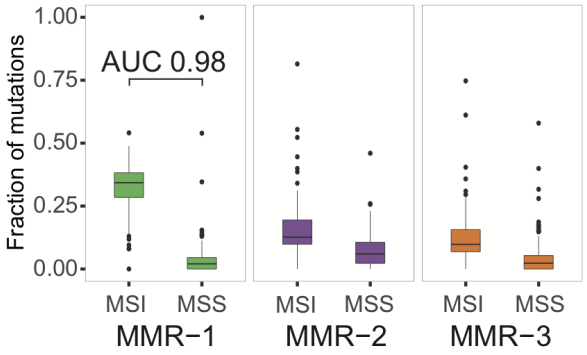


Figure 4.11: Relative contribution of MMR-1, MMR-2, and MMR-3 signatures to cancer samples clinically classified as MSI or MSS. Box plot with outliers shown as individual filled circles. Corresponds to Figure 3C in Meier et al. 2018.

By plotting a map of similarities between mutational spectra of all samples using a t-SNE representation (Maaten and Hinton 2008), we have observed some distinctive grouping of samples (Figure 4.12). A small yet well-defined cluster was formed by the samples whose mutational spectra were defined by POLE signature (brown). Samples dominated by Clock-2, 17-like or SNP signature tended to be located towards different ends of the map (blue, pink and grey, respectively).

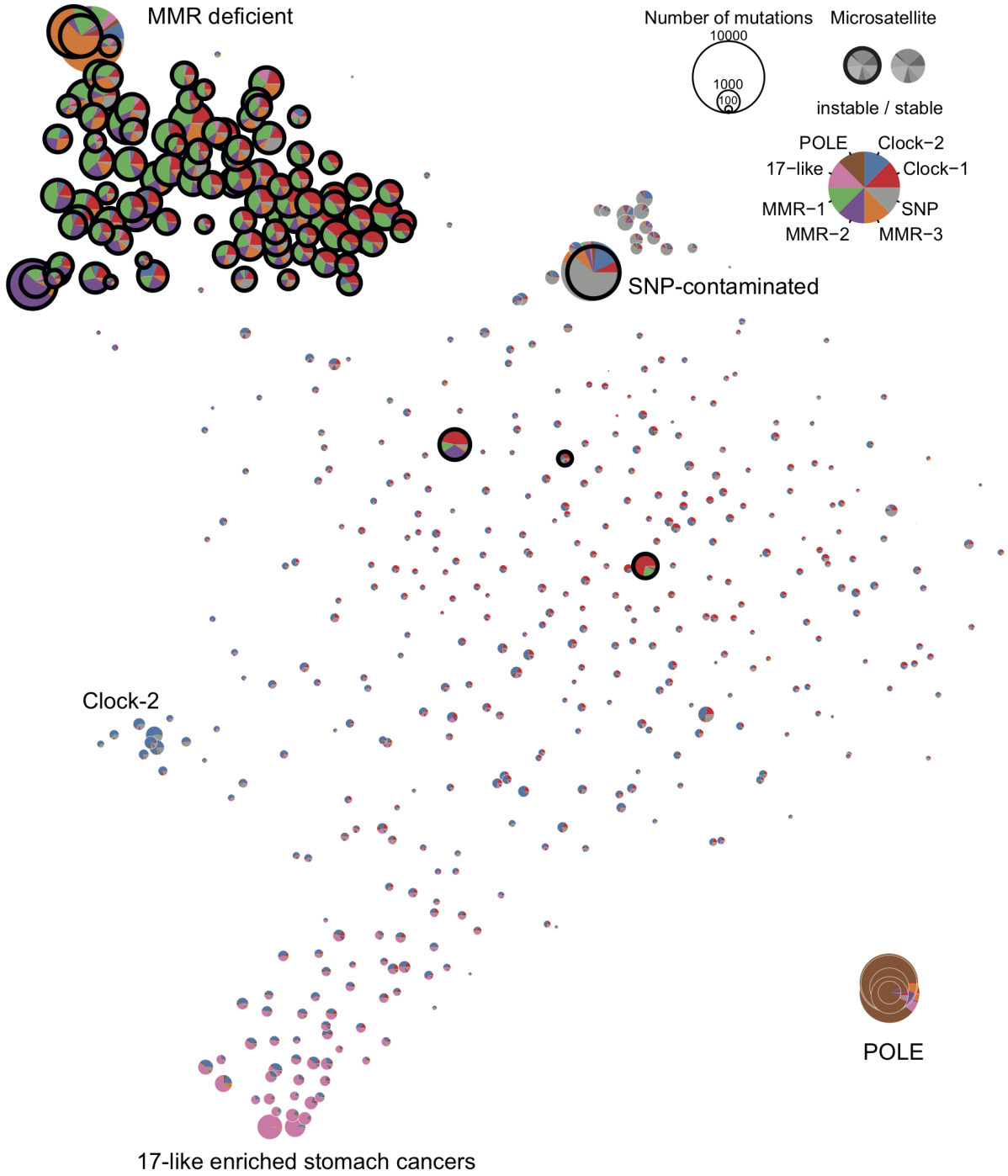


Figure 4.12: Two-dimensional representation of the mutational spectra composition across cancer samples. The size of each circle reflects the mutation burden. MSI samples are highlighted by a bold, black outline. The color of segments reflects the signature composition. Corresponds to Figure 3A in Meier et al. 2018.

The majority of MSI samples (circles with a black rim) concentrated in one cluster. These samples carried diverse combinations of signatures MMR-1-3 and Clock-1, but overall this group was mostly defined by the presence of the MMR-1 signature (green), while the MMR-2 (purple) and MMR-3 (orange) signatures occurred in a small number of tumours with high mutation burden. Moreover, the samples with the highest number of mutations were in fact largely described by a single signature, possibly reflecting the tendency of NMF to extract signatures from the most extreme cases first and then fit other samples as a linear combination of these basis vectors.

4.4.3 Mismatch repair and its interactions with other processes

Individual signatures often represent the most extreme ends of the mutational spectrum; a typical tumour, however, is usually represented by a linear combination of multiple processes. Given the different substrate specificities of MutS-α and MutS-β, MMR-2 and MMR-3 might reflect mutations arising by inactivation of unique subunits within these heterodimers. However, investigating MMR gene mutations and methylation status in these tumour samples, we observed few cases of MSH6, MSH3, PMS1 and MLH2 inactivation, which often occurred in combination with inactivation of other MMR genes, but did not correlate with the presence of any of the mutational signatures of interest.

In addition to the MMR-1-3 signatures, MSI samples were also partially composed by signature Clock-1. The Clock-1 signature closely resembled COSMIC signature 1, which was associated with 5-methyl-cytosine deamination and found to be correlated with the age at the time of diagnosis across a range of cancer types (Alexandrov et al. 2013b,Alexandrov et al. 2015). Given that there was no difference in the average age between MMR deficient and proficient groups of samples, this signature should be contributing a similar amount of mutations across these two groups. Since the average mutational burden in MMR deficient samples is higher, relative contribution of Clock-1 to these samples should be lower. However, we observed no change in the relative contribution, and a 10-fold increase (one-tailed t-test p-value $= 4.3 \cdot 10^{-22}$) in the absolute number of mutations assigned to Clock-1 in MMR deficient samples compared to proficient, similarly to signatures directly associated with MMR status (Figure 4.13).

This relationship indicates that MMR deficiency increased the rate of mutations resulting from spontaneous cytosine deamination. 5-methylcytosine can deaminate directly to thymine and form a G:T mismatch. A number of studies have previously suggested that mismatch repair can also detect and repair such mismatches as well as some alkylated and oxidised nucleotides outside of the context of replication (Bellacosa 2001, Tricarico et al. 2015b, Grin and Ishchenko 2016), performing so-called non-canonical mismatch

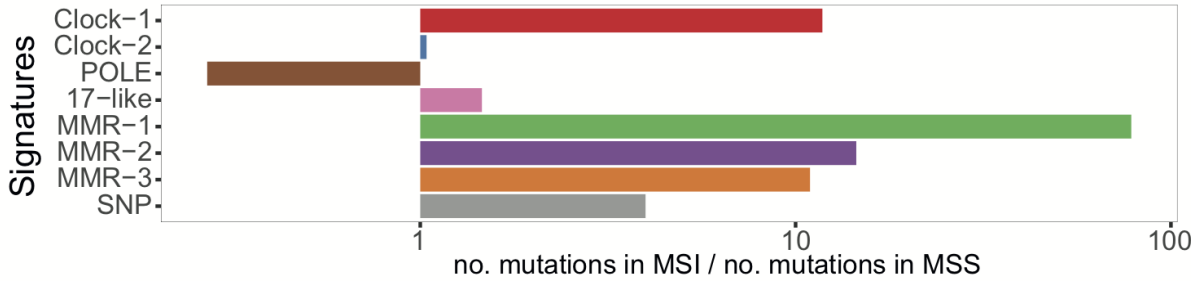


Figure 4.13: Fold-change in the average number of mutations assigned to different signatures in MSI samples compared to MSS samples. As expected, the number of POLE-related mutations is higher in MSS samples as all POLE-deficient tumours are MSS. Apart from MMR signatures, Clock-1 signature also contributes over 10 times more mutations to MSI samples than to MSS. Corresponds to Figure 3E in Meier et al. 2018.

repair. Typically, deamination-caused G:T mismatches in non-dividing cells would be processed by the base excision repair pathway. Notably, two of the COSMIC signatures associated with MMR deficiency, signatures 6 and 20, are very similar (cosine similarity 0.97) and only differ in the contribution of C>T mutations in NCG context, which may be reflecting a mixture between MMR mutational footprint and acceleration of cytosine deamination-caused transitions.

4.4.4 Deletions and insertions in repetitive sequences

The distribution of 1-bp indels extracted alongside with the base substitution profile for signature MMR-1 favours all of the bases equally (Figure 4.10), however in *C. elegans* we observed a huge shift towards indels of A and T. Mainly, this is due to a difference in homopolymer composition between the two systems: homopolymers in *C. elegans* genome mostly consist of poly-A and poly-T stretches (Figure 4.5), whereas the human exome features equal amounts of homopolymers formed by either base (Figure 4.14). In total, the human exome contained 976,390 homopolymer stretches between 4 and 43 bases in length, with A/T homopolymers contributing only about 55% or all homopolymers. For comparison, the *C. elegans* genome (with a length of 100 Mbps, approximately 3 times larger than the human exome) contained 3,433,785 homopolymers of which 94% were A/T stretches.

When adjusted for the difference in the available span of homopolymers, the frequencies of indels of A and T in *C. elegans mlh-1* and *pms-2* knockouts were still about 4 times higher than those of G and C (Figure 4.17). This indicates intrinsic differences in the processing of deletions and insertions of single bases.

Overall, 82% (25,093 out of 30,561) indels in the human dataset were single-base indels, and the majority of these occured in homopolymer runs: 69-72% of 1-bp indels/sample

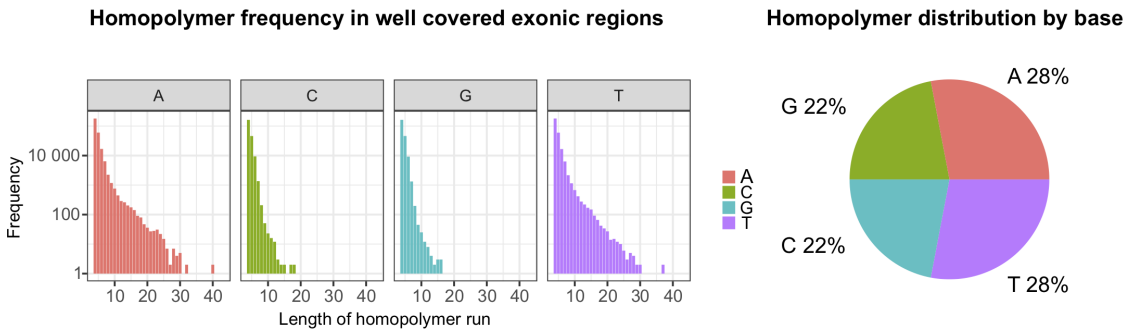
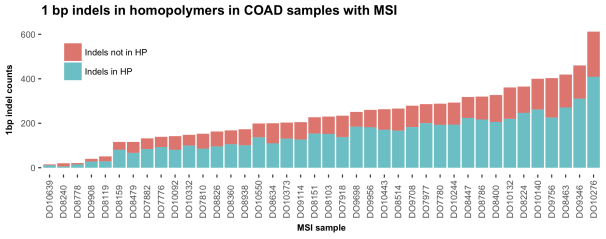


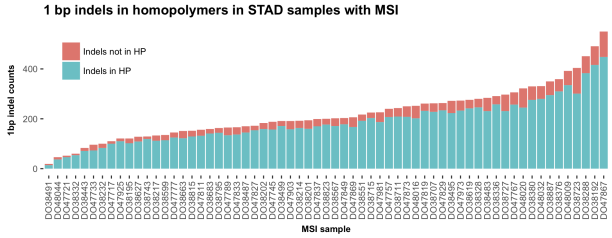
Figure 4.14: Distribution of homopolymeric sequences in the human exome. Corresponds to Supplementary Figure 4A in Meier et al. 2018.

in the COAD dataset and 91-93% 1-bp indels/sample for STAD (Figure 4.15).

The total amount of indels per homopolymers relative to the number of homopolymers in human data was much lower than that in *C. elegans*, hence, analysing the frequencies of indels per homopolymers of different length did not yield stable results for homopolymers longer than 9 bases.



(a)



(b)

Figure 4.15: Indels in (blue) and outside (red) of homopolymeric regions per sample in (a) COAD and (b) STAD dataset. Corresponds to Figure 4B in Meier et al. 2018.

4.5 *C. elegans* experimental MMR deficiency signatures correspond to signature MMR-1 in cancers

MMR mutant experiments in *C. elegans* allowed us to extract the mutational signatures of *mlh-1*, *pms-2* and *pole-4; pms-2* mutants. Having this data, we aimed to check how comparable these signatures are to the one active in cancer genomes. Using the trinucleotide frequency correction introduced in Chapter 2, we brought the *C. elegans* signatures into accordance with the human exome (using the counts from Rosenthal et al. 2016) (Figure 4.16). Indel fractions were adjusted using the ratios between the numbers of homopolymers built by respective bases in the *C. elegans* genome and in the human exome.

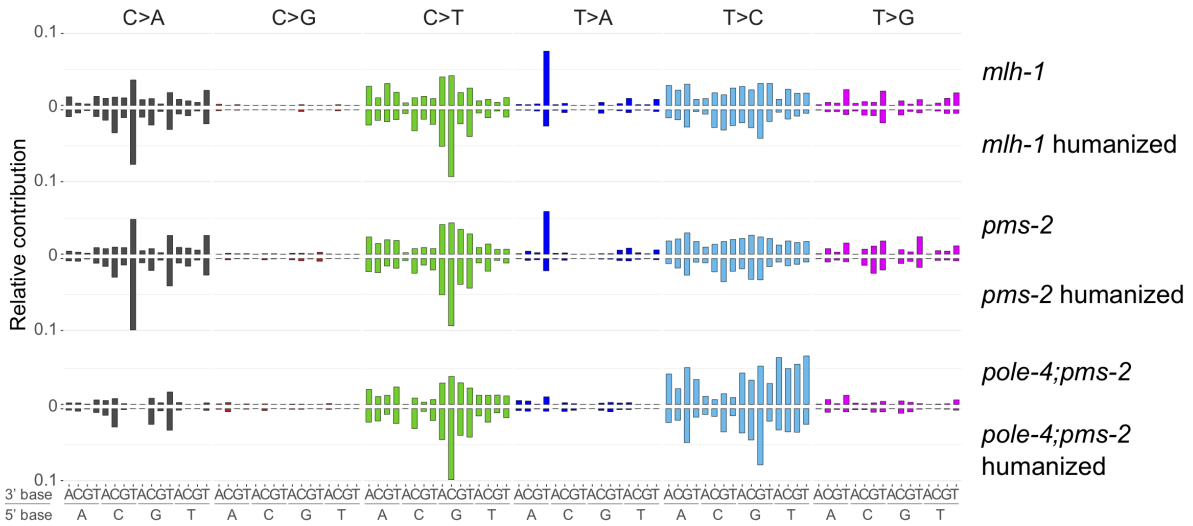


Figure 4.16: Original and humanised versions of *C. elegans* experimental signatures of *mlh-1*, *pms-2* and *pole-4; pms-2* knockouts. Corresponds to Figure 4A in Meier et al. 2018.

Of the three human MMRD-associated de novo signatures, only MMR-1 displayed similarity to the *C. elegans* MMR substitution patterns with cosine similarities of 0.84 and 0.81 to *pms-2* and *mlh-1* signatures, respectively (Table 4.1, Figure 4.17). Most of the discrepancy between the *C. elegans* MMR deficiency spectra and MMR-1 was coming from different levels of CpG>TpG mutations and change in frequency of ATT>AAT mutations.

C. elegans is lacking 5-methyl-cytosine (Greer et al. 2015). Hence, there is no spontaneous deamination, which explains much lower levels of C>T mutations at CpG sites. As 5-meC deamination appears to constitute a distinct mutational process, which is exacerbated by MMRD, it is likely that these mutation types define residual mutations

	mlh-1	*pms-2*	*pole-4; pms-2*
Clock 1	0.18	0.17	0.15
Clock 2	0.36	0.32	0.30
POLE	0.19	0.21	0.15
17-like	0.23	0.21	0.13
MMR-1	0.81	0.85	0.63
MMR-2	0.36	0.43	0.42
MMR-3	0.62	0.56	0.75
SNP	0.52	0.47	0.55

Table 4.1: Cosine similarity values for the comparison between humanised *C. elegans* derived MMR signatures and human de novo signatures (adjusted to human whole-exome trinucleotide frequencies).

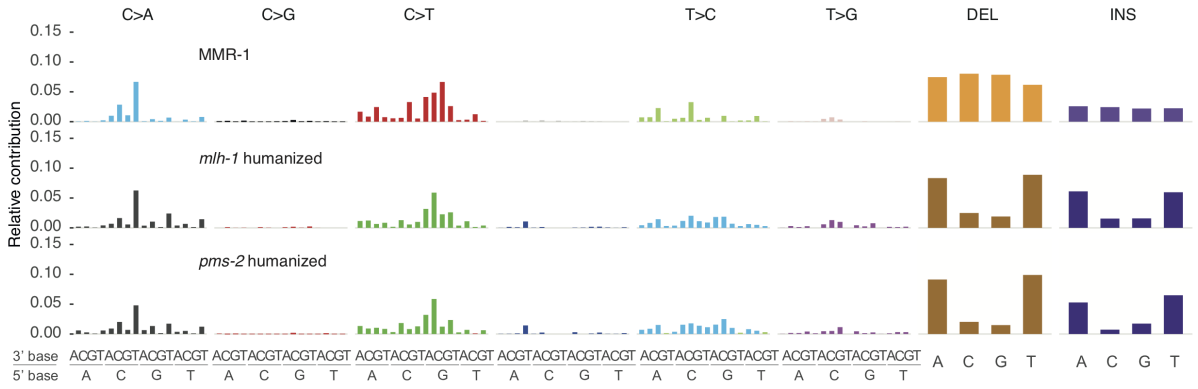


Figure 4.17: Signature MMR-1 and humanised mutational signatures of MMR deficiency in *C. elegans*.

rates which could not be attributed exactly by NMF. If these contexts were excluded, MMR-1 would show similarities of 0.92 to *pms-2* knockout signature, and 0.90 to *mlh-1* knockout signature. The human exome and the *C. elegans* genome also have a different composition of homopolymer types. The amount of poly-A and poly-T in the *C. elegans* genome is more than 6 times higher despite a comparable size (30 Mbps of the human exome, and 100 Mbps of the worm genome). Consequnently, the probability of more than one indel per homopolymer or of clashes between A-homopolymers and T-homopolymers is higher. Detection of indels at such positions is difficult since a read alignment with mismatches would be preferred over alignment with multiple gaps, and many indels in homopolymers are likely to be misinterpreted as a base substitution at the end of the respective homopolymer.

Having multiple experimental replicates allowed us to measure the variability of experimental signatures in *C. elegans*. In order to assess the variability of cancer-derived signa-

tures, we performed signature extraction for each jack-knife (leave-one-out) bootstrapped sample of the COAD/STAD dataset. Due to the presence of correlated processes (such as MMR deficiency and 5-meC deamination), computational signature decomposition may be unstable and produce different signatures. Hence, to quantify the corresponding change in similarity between the MMRD signatures, we calculated the cosine similarities between a random draw from 95% confidence intervals of *C. elegans* signatures and a randomly selected jack-knife draw from the MMR-1 signature (selected as the solution closest to the original MMR-1 signature in each subsample) (Figure 4.18).

Stability assessment of the similarity between *C. elegans* and human signatures shows that the similarity between humanised MMRD patterns and MMR-1 signatures varies substantially (inter-quantile range (IQR) of 0.66 to 0.76 for *mlh-1* and 0.70 to 0.80 for *pms-2*, respectively). Consistent with the assumption that dissection of MMRD and 5-meC deamination signatures may introduce additional variance, we observed that the range of similarity values shrinks as soon as we exclude C>T at CpG sites (IQR of 0.87-0.90 and 0.90-0.92, respectively) (Figure 4.18, red dashed line - the similarity for point estimates). These results support our conclusion about the signature MMR-1 being the closest reflection of the real MMRD signature consisting of mutations generated by the errors of replicative polymerases.

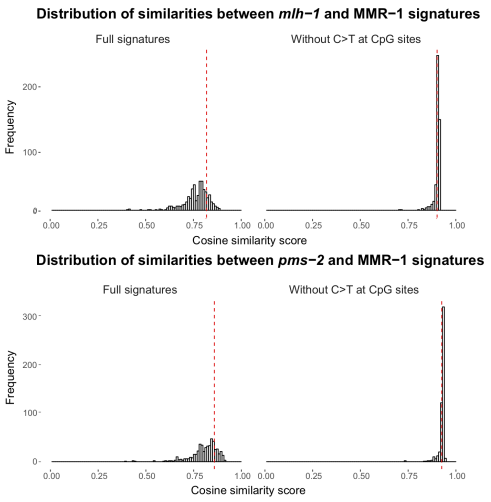


Figure 4.18: Distribution of similarities between humanised *mlh-1* signature (top) or *pms-2* signature (bottom) and human de novo signature MMR-1 with and without NCG>NTG mutations. Corresponds to Supplementary Figure 6 in Meier et al. 2018.

None of the human signatures showed notable similarity to the *pole-4; pms-2* mutation pattern. It may be because the dataset did not contain a sample with comparable concurrent defects of POLE4 and PMS2, or due to actual differences in the efficiency and proofreading abilities between human and *C. elegans* polymerase ε, however this question will require further investigation.

4.6 Discussion

In this chapter, I characterised the mutational signatures of mismatch repair deficiency across colorectal and stomach cancers and compared those to the mutational landscapes

associated with MutL MMR deficiency in *C. elegans*. MMR deficiency was of special interest as it led to the highest number of mutations and the most distinctive phenotype compared to other DNA repair deficiencies in our screen, surpassed only by that of the *pole-4*; *pms-2* double mutants, which exhibited 2-3 fold higher mutation rates. In addition to high substitution burden, MMR deficiency was associated with a high number of small indels in repetitive regions in both *C. elegans* and human cancers, consistent with the concept of polymerase slippage on homopolymer stretches, which should normally be repaired by the MMR system.

Out of the signatures we found to be associated with microsatellite instability in cancer cells, only one signature (MMR-1) was shown to be related to the experimental signatures of MMR deficiency found in *C. elegans mlh-1* and *pms-2* mutants. Taking into account the controlled nature of the *C. elegans* experiment, we postulated that MMR-1 might in fact reflect the underlying "basal" mutational process of DNA replication errors repaired by MMR conserved between species. Consistent with this, we find that MMR-1 showed high correlation with MSI status, and performed well in tumour classification (P-value $4.7 \cdot 10^{-55}$, AUC 0.985) into mismatch repair-deficient and proficient. Similarly, another study using human $MLH1^{-/-}$ organoid cells identified a mutational profile similar to that of MMR-1 (Drost et al. 2017).

The distribution of signature contributions indicated mixed origins of the other two MSI-associated signatures which tend to be concentrated in the most hypermutated samples. These signatures might be caused by a defect in another DNA repair gene, which led to an amplification of damage by replicative polymerases, or to a defect in replicative polymerase itself which contributed to the change of error profile, similar to *pole-4*; *pms-2* double mutants in *C. elegans* experiments or COSMIC signature 14, which was found to be associated with a concurrent loss of MMR and POLE proofreading ability (Haradhvala et al. 2018).

A particular value of comparison between *C. elegans* experiments and human data was coming from the absence of 5-cytosine methylation in *C. elegans*, which allowed us to extract an MMRD signature free from confounders. The most common MMRD signature found in cancer datasets, MMR-1, included an additional contribution of unrepaired 5-meC deaminations leading to C>T changes coming from the secondary role of MMR machinery in detecting and repairing the mismatches they cause. This part of human MMRD spectrum varies significantly depending on the age, tissue type and activity of other processes, which can cause ambiguities in unsupervised signature extraction. Any other human model would be prone to a similar issue. Due to the absence of such processes in *C. elegans* MMR knockouts, we could extract and compare the basic spectrum of mutations induced by MMR deficiency, which can be useful both for studying the error

profile of replicative polymerases and for diagnostics of MMR across individuals.

The similarity of the MMRD signatures across species also confirmed that the functionality and context-specific error rates of DNA polymerases ε and δ are well conserved between human and *C. elegans*. In this and previous chapters, I demonstrated how different DNA repair deficiencies are manifested in *C. elegans* genome, and confirmed the applicability of these insights to cancer data. The DNA repair deficiency, however, does not create damage on its own, the associated mutations have to result from another process which causes damage to the DNA. In the next chapter, I will consider a range of DNA damaging agents and study their mutational signatures in *C. elegans* and in humans.

Chapter 5

Experimental signatures of genotoxin exposures

5.1 Introduction

In the previous chapter, I demonstrated the ability of the *C. elegans* experimental screen to inform the mutational signature analysis in MMR deficient gastrointestinal cancers. DNA repair deficiencies were found to be a driving force in many cancers. Still, it is worth remembering that compromised DNA repair is not mutagenic by itself, but rather exacerbates the mutagenicity of endogenous or exogenous DNA damaging processes. In case of mismatch repair deficiency, it is the infidelity of DNA replication that causes most of the mutations observed in MMR deficient tumours. Many cancers have been associated with prolonged exposure to exogenous mutagens such as tobacco (Gandini et al. 2008), UV-light (Armstrong, Kricker, and English 1997) and ionising radiation (Shah, Sachs, and Wilson 2012, Health and Services 2016), which increases the probability of acquiring oncogenic mutations.

In tumours, exposures to genotoxins typically occur very early in the tumour development timeline (Stratton, Campbell, and Futreal 2009). Consequently, the mutations produced by these exposures are shared by a large number of tumour cells, and their spectrum creates a strong signal. The presence of a signature associated with a mutagenic exposure can serve as a molecular confirmation of the exposure, enhancing the epidemiological studies, or reveal previously unknown carcinogens (Helleday et al. 2008). This information can be important for cancer prevention: according to statistics on already identified carcinogens, lifestyle factors and infectious agents, about 40% of cancer could have been prevented by reducing or eliminating risk factors (Vineis and Wild 2014).

Apart from the potential causes, mutational signatures of genotoxins can also indicate

the origin of a tumour. Detection of cancer drug-associated signature in secondary tumours can help to distinguish between relapses and a *de novo* disease (Van Hoeck et al. 2019). Similarly, the presence of a signature associated with a mutagen that only affects one tissue or organ in a non-typical tumour – e.g. finding mutations induced by UV-light, which mostly affects the skin, in a lung tumour – strongly suggests a metastatic nature of such tumour (Liu et al. 2017b).

Studying genome-wide distributions of mutations induced by various genotoxins also provides insights into the mechanisms and specificities of their interaction with the DNA. Analysis of genome-wide signatures of mutations induced by different polycyclic aromatic hydrocarbons (PAHs) provided the basis to infer the diol-epoxide pathway of metabolic activation as the most likely (Kucab et al. 2019).

In this chapter, I will present a comprehensive analysis of mutation rates across different kinds of genotoxins, and consider the mechanistic details of mutation acquisition upon these exposures. To underline the significance of the signatures obtained in model systems to cancer research, I will compare the experimental signatures derived from *C. elegans* screen to mutational spectra observed in experiments conducted in different organisms and those detected in human cancers.

Contributions

As discussed in Chapter 2, the data was generated by Bettina Meier and colleagues from Anton Gartner's group at the University of Dundee. All downstream bioinformatics analyses were performed by me. A brief overview of the signatures of genotoxins and their comparison to mutational signatures of genotoxins in human cells and cancer was submitted for publication as a part of Volkova et al. 2019.

5.2 Experimental signatures of mutagenesis upon genotoxin exposure

To enrich the understanding of mutagenesis induced by different kinds of genotoxins, we exposed both wild-type and DNA repair-deficient *C. elegans* to 12 mutagenic agents. The genotoxins used in the screen were chosen such that a wide range of DNA damage could be observed. The list of mutagens we studied includes irradiations, alkylating agents, crosslinkers and agents inducing bulky adducts.

In particular, we investigated the mutational signatures of UV-light by using a source of simulated UV-B radiation, as well as the spectra resulting from exposures to aflatoxin-B1, a naturally occurring mycotoxin produced by certain species of fungi capable of inducing

liver cancers (Kew 2013), and aristolochic acid – a phytochemical associated with kidney damage and elevated rates of urothelial cancers (Arlt, Stiborova, and Schmeiser 2002). We included several genotoxins used for chemotherapy treatment such as cisplatin, hydroxyurea (also known as hydroxycarbamide), Mitomycin C, and both γ- and X-rays to study the mutational properties of ionising radiation. In addition, we enhanced the screen using three mutagenic substances commonly used in genetics: monofunctional alkylating agents ethyl methanesulfonate (EMS), dimethyl sulfate (DMS) and methyl methanesulfonate (MMS).

Apart from inducing a diverse and robust range of mutations, the mutagen exposure should still allow producing viable progeny, which will be used to collect DNA for sequencing. Hence, the exposure dose had to be adjusted not to cause severe mortality in *C. elegans*. Consequently, the genomic DNA of the adult progeny derived from single fertilised eggs was sequenced. To ensure that the progeny was able to reach the three-day adulthood stage, the highest dose for each exposure experiment was chosen such that no more 25% reduction in the viability of embryos occurred.

5.2.1 Signatures of alkylating agents

In its simplest form, DNA adducts are methyl or ethyl groups, and three common agents of alkylation are methyl methanesulfonate (MMS) and dimethylsulfonate (DMS), both leading to methylation, and ethyl methanesulfonate (EMS), leading to ethylation of DNA bases. Generally, these alkylating agents are not considered to be environmental contaminants. However, there have been reports of accidental EMS contamination of Viracept – film-coated tablets used as an antiviral medication (Gerber and Toelle 2009), and evidence of occupational DMS exposure in the chemical industry, where the reagent is being used to alkylate organic substrates (Rippey and Stallwood 2005).

EMS

The substance that yielded the highest number of mutations was EMS – a monofunctional alkylating agent with a chemical formula $C_3H_8SO_3$ (Figure 5.1a). EMS is one of the most commonly employed mutagens for randomly introducing mutations into genomes during genetic screens. It was shown to be mutagenic in many systems from viruses to mammalian cells, including T4 bacteriophages, *D. melanogaster*, and *C. elegans*, where it was used to perform unbiased genetic screens (Hartman et al. 2014, Huang 1981, Drabløs et al. 2004).

EMS mostly alkylates guanines at the O6 position leading to the formation of O6-alkylguanines (Figure 5.1b). Due to an alkyl group on the oxygen atom, O6-alkylguanine

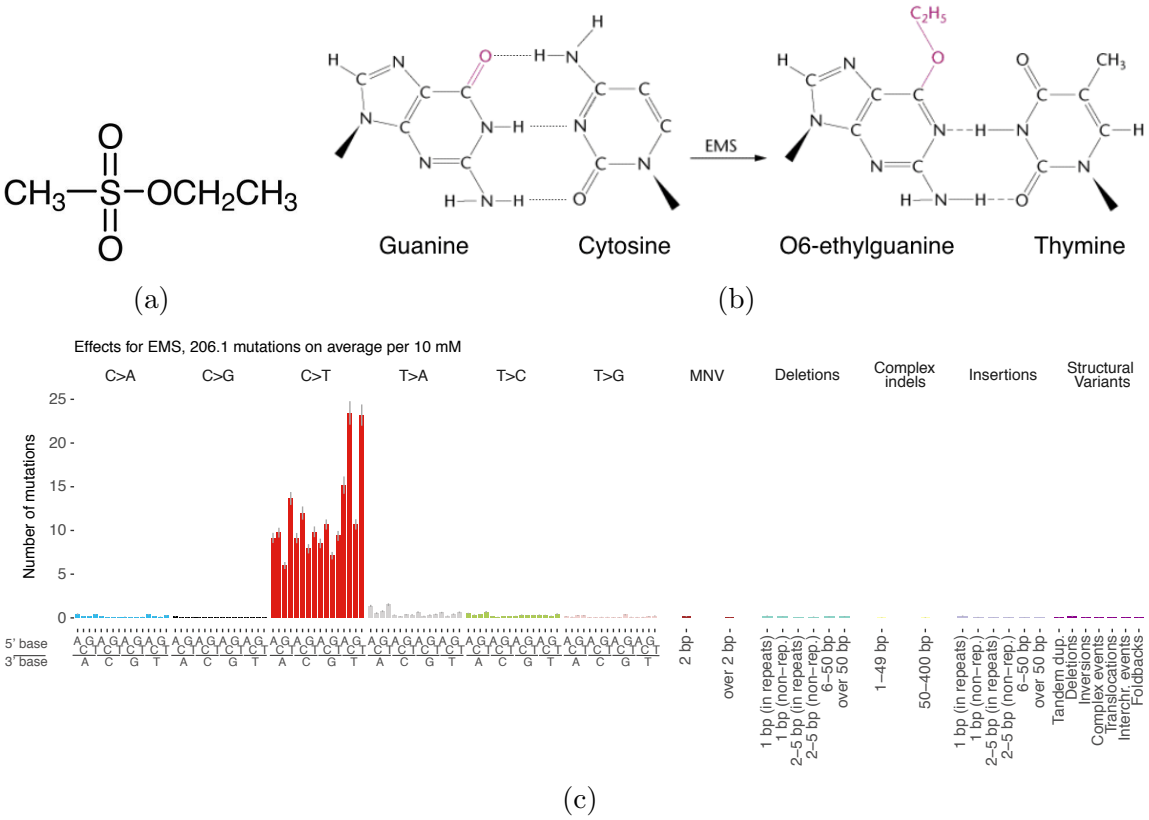


Figure 5.1: EMS mutagenesis. (a) Chemical structure of EMS molecule. (b) Modification of guanine upon EMS exposure turns it into a O6-ethylguanine, which can pair with thymine, leading to G:C>A:T mutations. Adapted from Klug and Cummings 2006. (c) Experimental mutational signature of EMS in *C. elegans*.

does not form a triple bond with cytosine and is prone to be paired with thymine, which can result in C:G>T:A transition (Brookes and Lawley 1961). Many of these mutations, when occurring in coding regions, generate premature stop codons, leading to strong loss-of-function mutations (Flibotte et al. 2010). This base modification was previously described to account for 2-3% of ethylation upon EMS treatment. O6-alkylguanines are typically repaired by direct repair enzymes, namely the O6-alkylguanine alkyltransferases (AGT1 in *C. elegans* or MGMT in humans), which remove the alkyl group without altering the base.

The rest of the damage incurred by EMS are non-mutagenic lesions such as N7-alkylguanine, which does not directly lead to mutagenesis as it does not affect guanine to cytosine base pairing. However, N7-alkylation makes guanine more prone to depurination, creating additional AP-sites (Boysen et al. 2009).

In agreement with other model systems, EMS mostly caused base substitutions in wild-type *C. elegans*. On average, exposure with 10 mM of EMS led to about 230 mutations per genome in the progeny, most of which were C>T transitions with a small preference

for TpCpC and TpCpT contexts (Figure 5.1c).

MMS and DMS

In contrast to EMS, exposure to alkylating agents MMS and DMS yielded a different mutational spectrum. MMS is a monofunctional alkylating agent with a chemical formula $C_2H_6O_3S$ (Figure 5.2a). While having a different chemical structure (Figure 5.2a, $C_2H_6O_4S$), DMS causes the same kinds of DNA damage as MMS (Figure 5.2b). DMS is a highly mutagenic methylating agent widely used in the chemical industry.

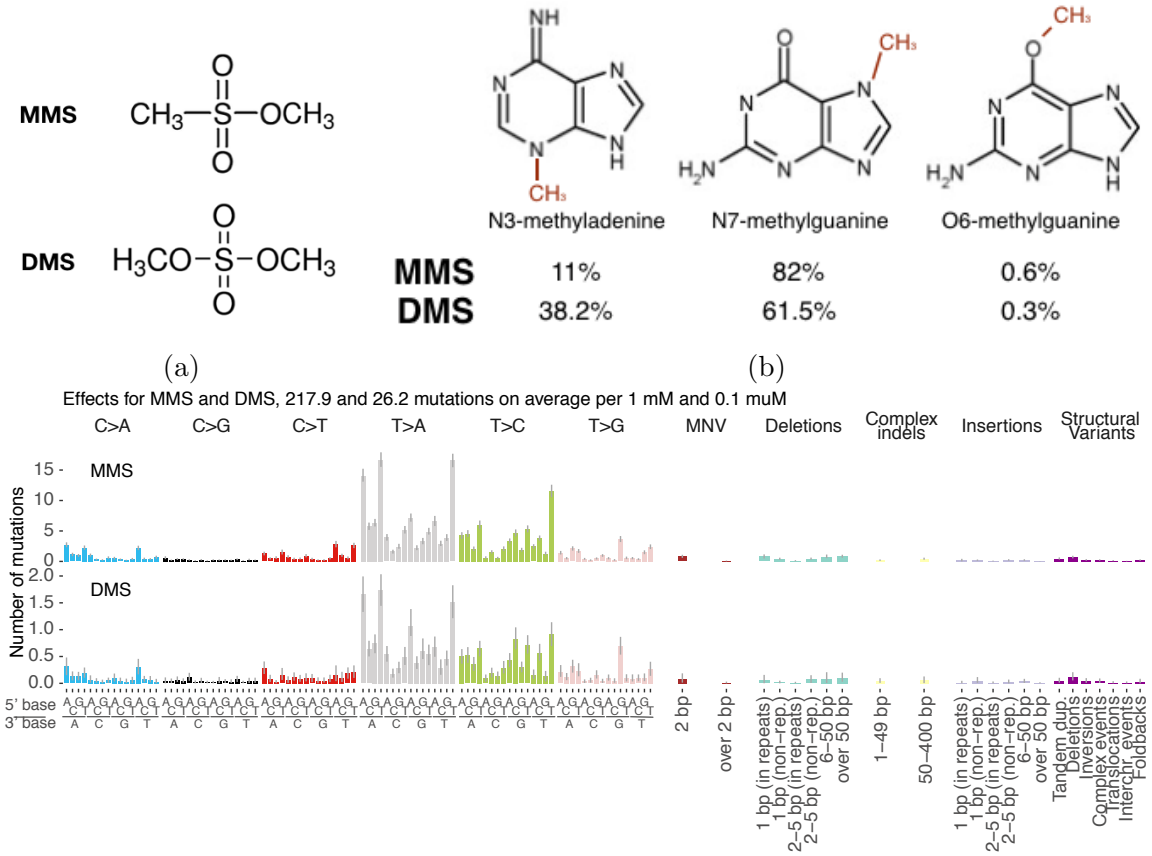


Figure 5.2: DMS and MMS mutagenesis. (a) Chemical structure of MMS (top) and DMS (bottom) molecule. (b) Most common modifications of DNA bases upon DMS/MMS exposure: N3-methyladenine, N7-methylguanine, O6-methylguanine. Adapted from Chadt et al. 2008. (c) Experimental mutational signatures of MMS and DMS in *C. elegans*.

Both MMS and DMS affect DNA by methylating their bases. Up to 82% of MMS-induced adducts are expected to be N7-meG (which is not mutagenic), accompanied by 11% of N3-meA, which lead to T>N mutations and a small amount (0.6%) of O6-meG (Wyatt and Pittman 2006). The rest of damage is comprised of less impactful adducts

(N1-meA, N7-meA, N1-meG, N3-meG, N3-meC, O2-meC, N3-meT, O2-meT, and O4-meT). The distribution of DMS damage is similar but not identical: it typically introduces N7-meG in about 61.5% of cases, N3-meA in 38.2%, and O6-meG in the remaining 0.3% of cases (Chadt et al. 2008).

In *C. elegans*, both agents caused nearly identical mutation spectra (with cosine similarity of 0.98) characterised by T>A and T>C substitutions (Figure 5.2c). Evidence suggests that the T>A and T>C changes might result from an error-prone bypass of N3-methyladenine, a lesion associated with stalling of replicative polymerases (Fronza and Gold 2004). As described above, the rare alkylation of O6 in guanines is equally mutagenic due to O6-methylguanine being able to mispair with thymidine (Brookes and Lawley 1961).

Interestingly, the ratio of the damage probabilities is different from the observed mutation ratio: for both MMS and DMS, 85% of the damage consisted of T>N changes, whereas based on the reported frequencies of the meA and meG adducts, it should have been 95% for MMS and over 99% for DMS. This indicates a difference in repair efficacy of different bases: seemingly, the probability of correctly bypassing N3-meA is higher than the probability of repairing O6-meG before it causes a C>T mutation during replication.

On average, MMS caused 218 mutations per 1 mM per genome, while DMS demonstrated a lower rate of only 26 mutations per median unit dose 0.1 μM (Figure 5.2c), implying a different reactivity of introducing DNA methylation, lower rates of cellular uptake, or higher rate of metabolism.

5.2.2 Agents introducing bulky DNA adducts

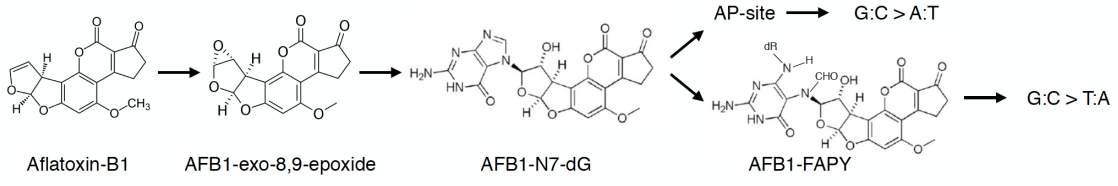
Various genotoxic substances can generate highly reactive intermediates upon metabolisation. These molecules are capable of covalently bonding to the DNA resulting in bulky adducts to the nucleotides, which can distort the spatial conformation of the double helix. Such damage is typically repaired by the nucleotide excision repair system, and unrepaired adducts can lead to base substitutions as a result of error-prone bypass, but also replication stalling (Minca and Kowalski 2010).

Aflatoxin B1

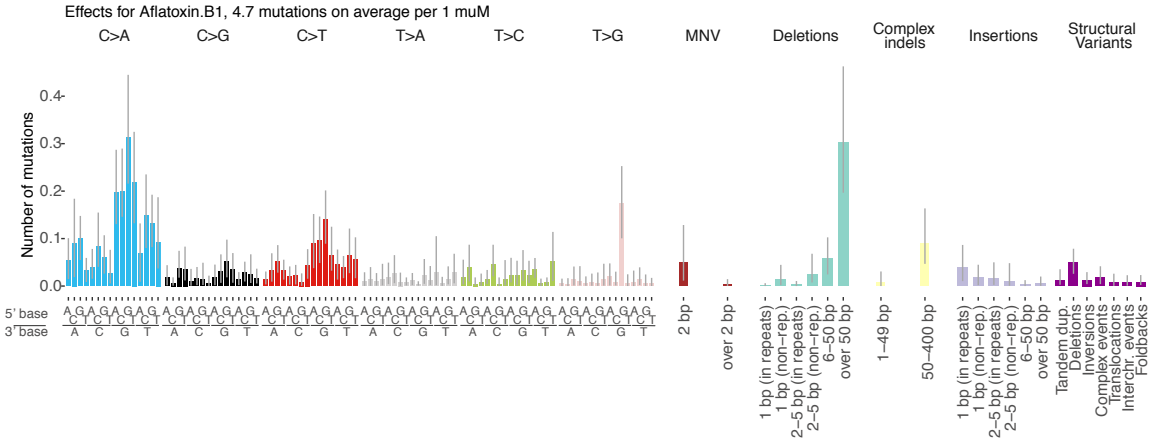
Aflatoxin B1, a mycotoxin produced by a fungus *Aspergillus flavus*, has been associated with the development of liver cancers in populations such as Sub-Saharan and South African, or South-East Asian where the food is often stored in hot and humid conditions and may be contaminated with aflatoxin (Kew 2013).

Aflatoxin B1 requires metabolic activation: an active form of this substance, exo-

AFB1-8,9-epoxide, can bind to the N7 position of guanine residues forming 8,9-dihydro-8-(N7-guanyl)-9-hydroxyaflatoxin-B1 guanosine adducts (AFB1-N7-dG, Figure 5.3a). If left unrepaired, these adducts can undergo spontaneous depurination, which would usually cause a G:C>A:T change if unrepaired, or rearrange to an open-ring AFB1-formamidopyrimidine (AFB1-FAPY) adduct leading to G:C>T:A transversions upon replication due to error-prone bypass by polymerase ζ (Lin et al. 2014), each causing characteristic peaks in the base substitution spectrum (Figure 5.3b). If the translesion synthesis is delayed, a stalled replication fork can result in single- or double-strand breaks, generating medium-size deletions via polymerase θ mediated end-joining (Roerink, Schendel, and Tijsterman 2014, Larsen et al. 2017).



(a)



(b)

Figure 5.3: Aflatoxin-related mutagenesis. (a) A scheme of aflatoxin modification: aflatoxin is metabolised to epoxide, which creates AFB1-N7-dG lesions; they can further undergo depurination and cause a G>A change, or transform into AFB1-FAPY causing a G>T mutation. (b) Experimental mutational signatures of aflatoxin B1 in *C. elegans*.

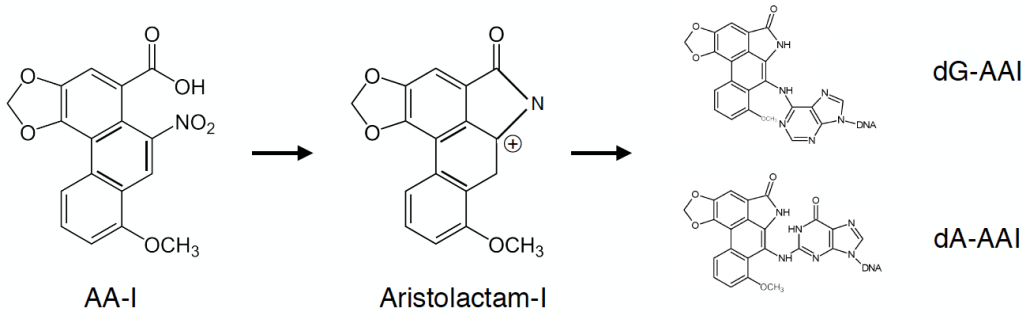
Aflatoxin has been shown to cause similar kinds of DNA damage and mutations in *C. elegans* cells as well as other model systems (Leung et al. 2010, Meier et al. 2014, Huang et al. 2017a), indicating conserved metabolic activation and repair patterns in humans, mice and *C. elegans*. Bulky adducts created by aflatoxin-B1 led to a mutational spectrum consisting mostly of C>A and C>T mutations in agreement with its adducts being concentrated on guanines (Figure 5.3b), with a small fraction of 50-400 bp long

deletions. Treatment with 1 μM of aflatoxin caused on average 4 mutations in wild-type *C. elegans*.

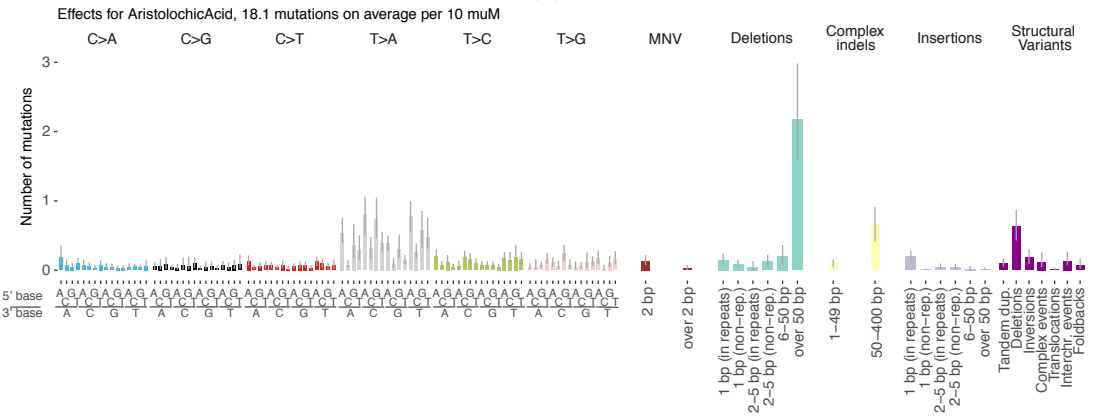
Some properties of aflatoxin mutagenesis across DNA repair mutants were already studied in Meier et al. 2014.

Aristolochic acid

Aristolochic acid (AA) is a natural nitro-compound associated with Balkan Endemic Nephropathy, a chronic renal disease frequently leading to upper urothelial cancer (Arlt, Stiborova, and Schmeiser 2002, Poon et al. 2015), and also hepatocellular carcinomas (Ng et al. 2017). AAs, the most abundant of which is AAI (Figure 5.4a), are phytochemicals produced by *Aristolochia clematitis*, the European birthwort, a weed growing in wheat fields leading to the contamination of baking flour. Other members of *Aristolochia* family are commonly used in traditional Chinese medicine (Tian-Shung et al. 2005).



(a)



(b)

Figure 5.4: Aristolochic acid-induced mutagenesis. (a) Chemical structure of AA-I and its reactive metabolite aristolactam, which can cause adducts to purines. Adapted from Sidorenko et al. 2012. (b) Experimental mutational signatures of aristolochic acid I in *C. elegans*.

Upon metabolisation, AA is enzymatically reduced to form several carcinogenic in-

termediates. Among others, it produces aristolactam which forms adducts on the exocyclic amino groups of deoxyguanosine and deoxyadenosine, dG-AAI and dA-AAI (Pfau, Schmeiser, and Wiessler 1990, Figure 5.4a). Of these, adenine adducts are 5 to 10 times more abundant than the guanine ones and also more mutagenic: both of them mostly lead to a misincorporation of an adenine (Attaluri et al. 2009). dA-AA causes A:T>T:A changes, which are usually repaired by TC-NER and are more persistent in untranscribed regions (Sidorenko et al. 2012). This profile was first measured by base changes in the p53 locus of urothelial cancers associated with AA and more recently was confirmed by next-generation sequencing on a genome-wide scale (Poon et al. 2013, Poon et al. 2015).

Mutational signature of aristolochic acid (AA) exposure in *C. elegans* showed a characteristic T>A substitution spectrum with a low amount of C>N changes, which was about 5 times lower than the amount of T>N substitutions in agreement with the reported frequencies of dA and dG adducts, indicating similar mutagenicity of each lesion. In addition, we observed a high prevalence of 50-400 bp long deletions and complex indels (Figure 5.4b), which are likely to arise due to MMEJ repair of double-strand breaks (DSBs) caused by a delayed or failed translesion synthesis similar to aflatoxin-induced mutation spectra. On average, exposure with 10 μM of AA led to approximately 10 mutations per *C. elegans* genome.

5.2.3 Crosslinking agents

Cisplatin

Cisplatin (Figure 5.5a) is a widely used chemotherapy drug, mostly due to its crosslinking properties, which stall replication and lead to apoptosis of cycling cells in which the damage cannot be repaired sufficiently fast. Cisplatin molecules can create intra- and inter-strand crosslinks between N7 positions of purines, as well as platinum mono-adducts or DNA-protein crosslinks (Figure 5.5b). Over 90% of the lesions induced by cisplatin are intra-strand crosslinks: on average, two-thirds of them are GpG crosslinks, one-quarter – ApG lesions, and the rest are GpNpG crosslinks (Cohen and Lippard 2001).

Intra-strand crosslinks and DNA-protein crosslinks are usually repaired by the NER and Fanconi Anaemia pathways, whereas monoadducts can also be repaired by BER. Hence, most of the mutagenesis induced by genotoxin is caused by polymerase ζ-dependent translesion synthesis over the intrastrand crosslinks, which leads to single-base and dinucleotide substitutions (Noll, Mason, and Miller 2006).

In *C. elegans*, cisplatin exposure induced C>A transversions in a CpCpC and CpCpG context as well as deletions and structural variants (Figure 5.5b, Boot et al. 2018, Meier and Gartner 2014). A high number of dinucleotide substitutions at the sites of intrastrand

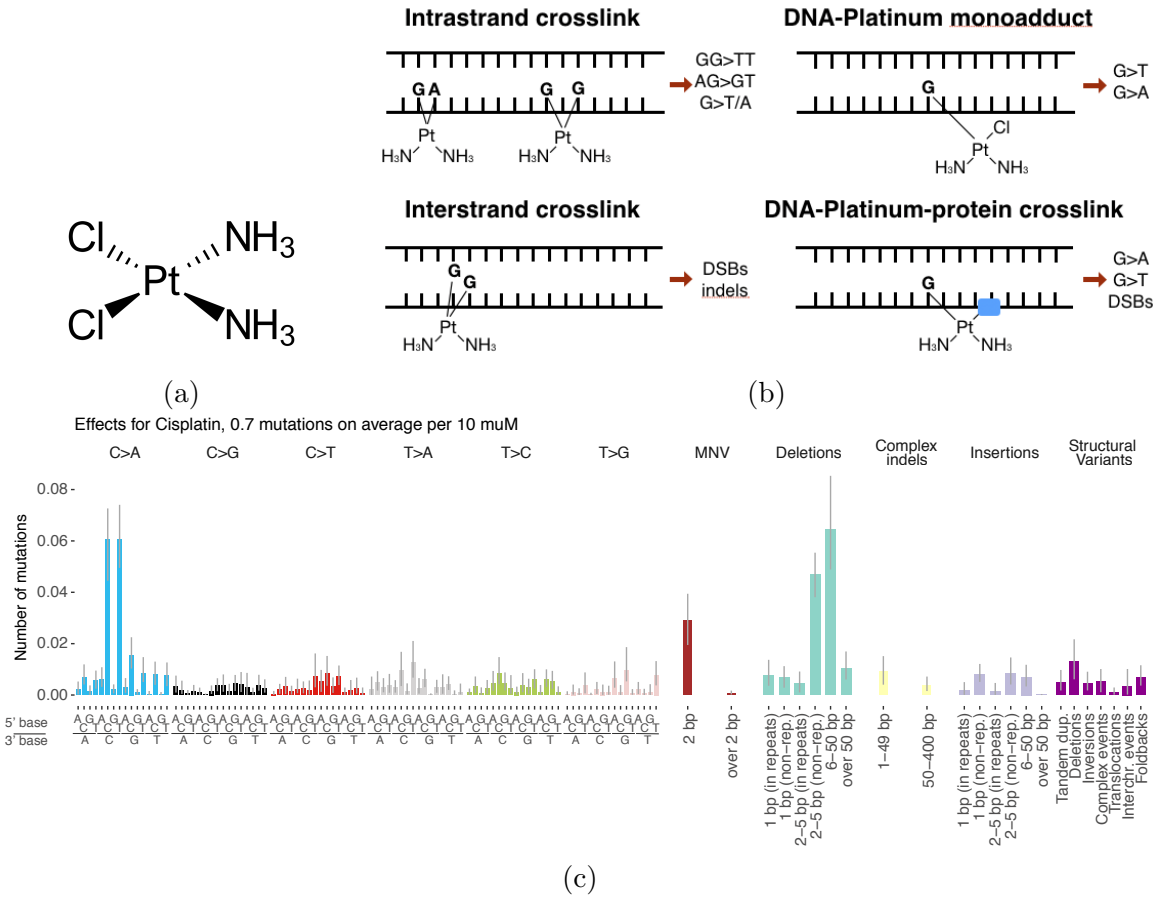


Figure 5.5: Cisplatin mutagenesis. (a) Chemical structure of cisplatin molecule. (b) Damage types caused by cisplatin. Adapted from Rabik and Dolan 2007. (c) Experimental mutational signature of cisplatin in *C. elegans*.

crosslinks were observed, mostly of CT>AC and TG>GT types (Figure 5.6), consistent with previous reports (Meier et al. 2014). In cisplatin-treated cancer cell lines, equal amounts of CC>AD (where D stands for A, G or T) and CT>AM (M = A,C) dinucleotide substitutions were observed (Boot et al. 2018), potentially indicating different repair capacities for GpG and ApG crosslinks in *C. elegans* and human cell lines. However, some reports suggest that CC>AA (GG>TT) mutations are wide-spread across cancers, correlate with the age of diagnosis and can arise in normal human cells (Alexandrov et al. 2018). Interestingly, the computationally derived DNV signature found to be associated with treatment with platinum drugs in cancers was on two-thirds comprised of CT>AA, and one-third of CT>AC, indicating that when the damage occurs in the same dinucleotide, the repair can be different (Alexandrov et al. 2018).

In addition, we observed a high proportion of deletions between 2 and 50 bp in size and occasional large deletions in the mutational spectrum generated by cisplatin exposure in *C. elegans* (Figure 5.5c). These are likely to stem from the homology-directed repair

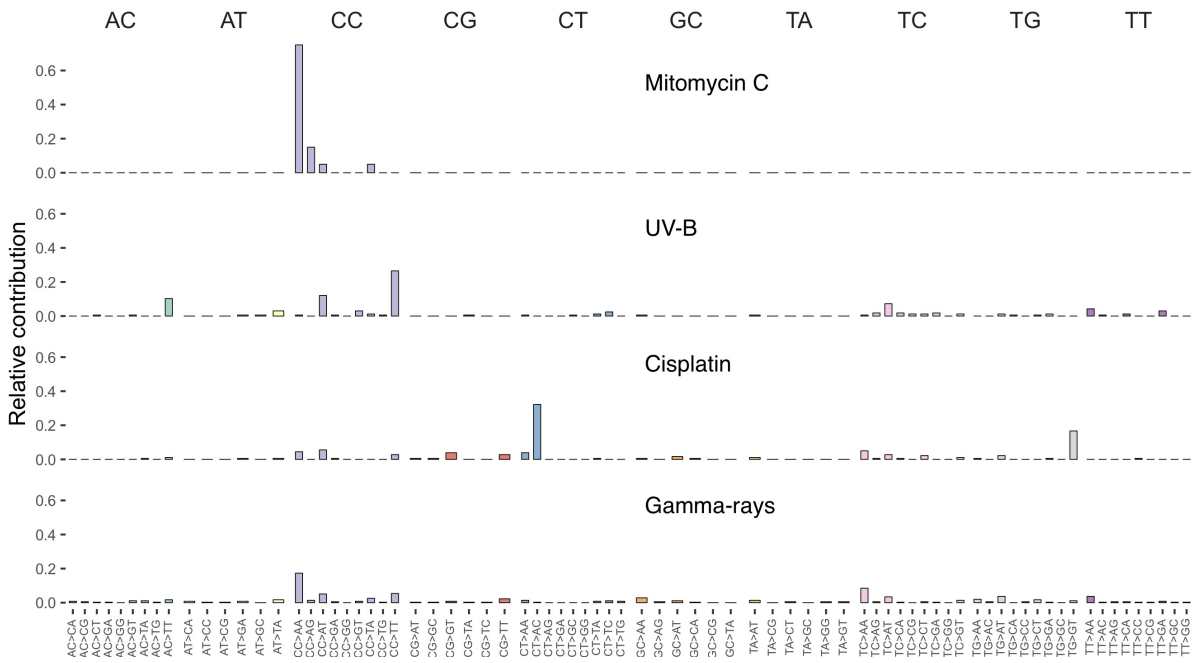


Figure 5.6: Distributions of dinucleotide substitutions for mitomycin C, UV, cisplatin and radiation exposures.

of interstrand crosslinks (Jonnalagadda, Matsuguchi, and Engelward 2005).

Mechlorethamine

Mechlorethamine, or chlormethine, is an analogue of mustard gas. It is a bi-functional alkylating agent capable of introducing interstrand DNA crosslinks by reacting with N7 end of two guanines from opposite strands. Due to its crosslinking abilities, mechlorethamine is used as a chemotherapy drug for the treatment of Hodgkin lymphomas (Engert, Wolf, and Diehl 1999).

Similar to cisplatin, mechlorethamine (Figure 5.7a, $C_5H_{11}C_l2N$) can create crosslinks, especially interstrand crosslinks between guanines in CpGpN and GpCpN contexts, and monofunctional adenine and guanine adducts (Povirk and Shuker 1994). Nucleotide excision repair efficiently removes the adducts, but is not as efficient in resolving interstrand crosslinks: if not repaired, interstrand crosslinks will be repaired by homologous recombination repair which can result in a loss of genetic information (Jonnalagadda, Matsuguchi, and Engelward 2005).

Consistent with previous reports (Wijen, Nivard, and Vogel 2000, Meier et al. 2014), 10 mM of mechlorethamine caused on average 15 mutations in wild-type *C. elegans*, most of which were small deletions between 2 and 50 bp in length (Figure 5.7c). Similar to cisplatin, they can be caused by the homology-directed repair of interstrand crosslinks, which was shown to have a higher chance of deletions and insertions compared to the HR

repair of spontaneous DSBs (Jonnalagadda, Matsuguchi, and Engelward 2005).

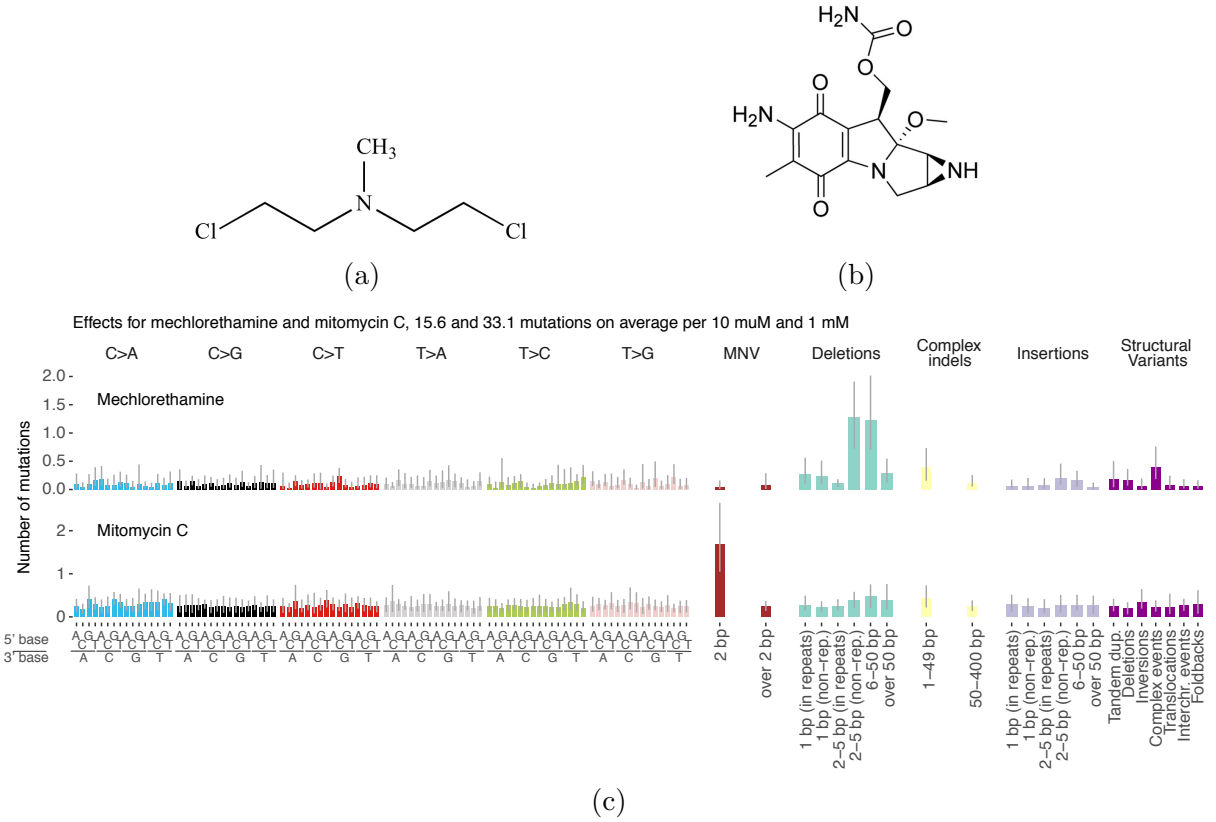


Figure 5.7: Mechlorethamine and mitomycin mutagenesis. (a) Chemical structure of mechlorethamine molecule. (b) Chemical structure of mitomycin C molecule. (c) Experimental mutational signatures of mechlorethamine and mytomycin C in *C. elegans*.

Mitomycin C

Another crosslinking drug used in chemotherapy, mitomycin C, demonstrated a mutational spectrum almost exclusively defined by dinucleotide substitutions (Figure 5.7c). Its main mechanism of action mostly introduces N7-guanine adducts and crosslinks between guanine residues of 5'-CpG-3' sequences through the minor groove of DNA. Previously, mitomycin C was shown to cause deletions containing CpG dinucleotides in *C. elegans* (Tam, Chu, and Rose 2016), however, our experiments did not yield a high number of deletions.

Crosslinks between the neighbouring guanines which are on the same strand are prone to result in dinucleotide substitutions due to error-prone bypass during replication (Noll, Mason, and Miller 2006). In wild-type *C. elegans*, 1 mM of mitomycin C yielded about 33 mutations (Figure 5.7c), with a high prevalence of CC>AA and CC>AG changes (Figure 5.6).

5.2.4 Electomagnetic radiation

Highly energetic electromagnetic radiation interacts with the atoms in the DNA and other molecules in the cell, introducing DNA damage and inducing oxidative damage. The severity and proportions of damage caused by electromagnetic waves vary depending on the wavelength. Therefore UV (λ = 10-400 nm), X- (0.01-10 nm) and gamma ($\lambda <$ 0.01 nm) are potent mutagens and known carcinogens.

UV

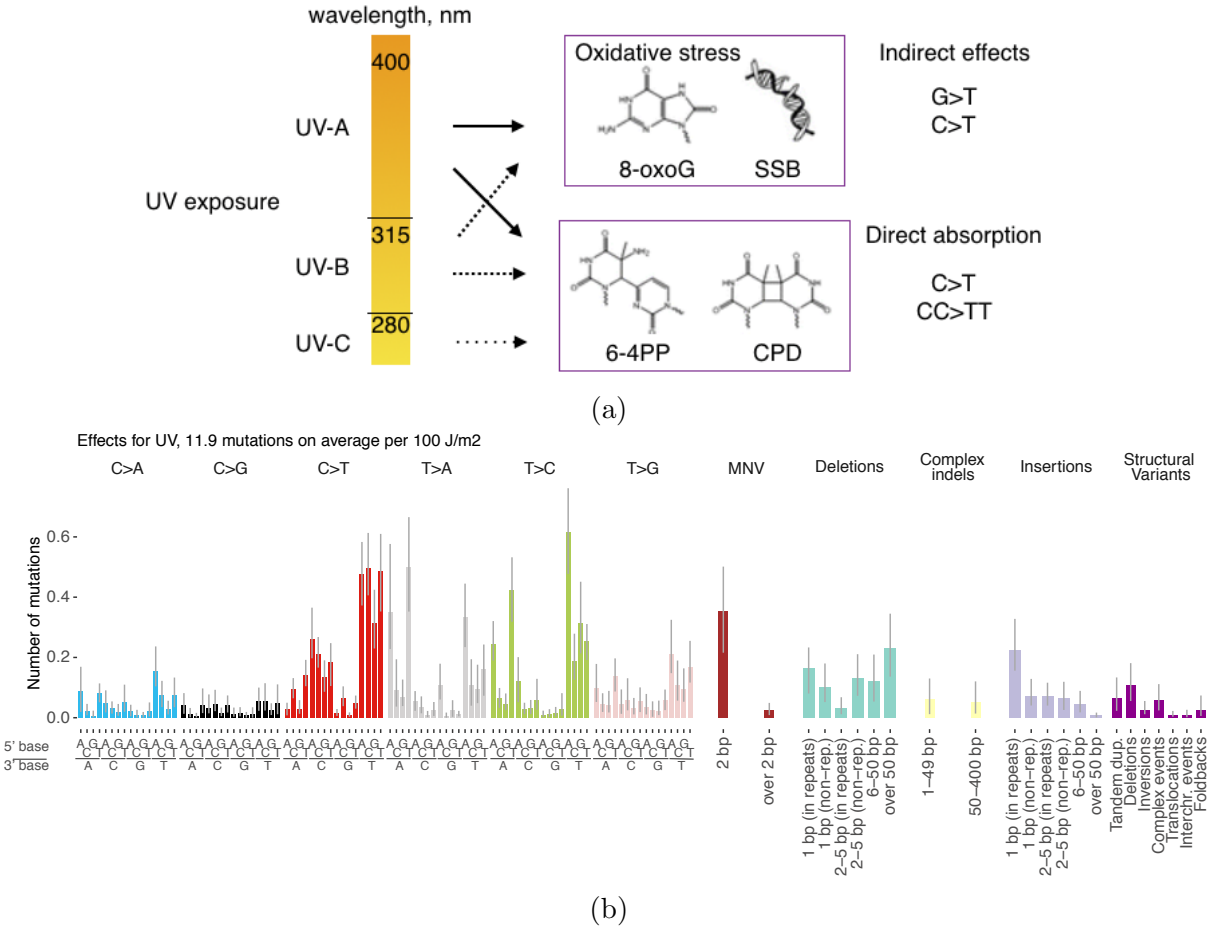


Figure 5.8: UV mutagenesis. (a) Types of UV exposures and DNA damage. (b) Experimental mutational signatures of simulated UV-B radiation in *C. elegans*.

Sunlight exposure is one of the most common cancer risk factors. About 3% of the solar irradiation at ground level covers the UV spectrum ranging from 400 nm (UVA) over 315 nm (UVB) to 280-100 nm (UVC; Figure 5.8a). Interaction of UV-light with DNA causes direct damage consisting of pyrimidine dimers (CPDs) and 6-4 pyrimidone photoproducts (6-4PPs). Whole-genome assessment of UV damage distribution in yeast

reported over 70% of CPDs occurring at TpT dinucleotides, followed by TpC and CpT (about 12% each), with only 5% occurring at CpCs, whereas 6-4PPs were, in contrast, enriched at CpTs, with the rest almost equally divided between TpT and CpC (Bryan et al. 2014). In humans, CPDs occur 3 to 5 times more frequently than 6-4PPs (Cadet and Douki 2018).

Typically, the lesions classified as direct damage are repaired by nucleotide excision repair, or replicated over by translesion synthesis polymerases. Polymerases η can bypass CPDs and tends to insert an adenine opposite a damaged base (Matsuda et al. 2001), thus generating the characteristic C>T and CC>TT mutations while correctly bypassing the most abundant TpT dimers (Ikehata and Ono 2011). 6-4PP is a more helix-distorting lesion, and requires a joint effort of two or more TLS polymerases: polymerases η or ι can insert the first base, which for pol η is often a G opposite a 3' T (Johnson et al. 2001), and polymerase ζ extends to the second base (Yoon, Prakash, and Prakash 2010), which can potentially generate a T>C mutation.

Additionally, UV can cause indirect damage (or 'dark' UV damage) by inducing reactive oxygen species, single-strand breaks (SSBs) and depurination of bases leading to abasic sites (Rastogi et al. 2010, Figure 5.8a).

Exposure to simulated UV-B radiation in *C. elegans* experiments showed characteristic C>T transitions in a YpTpH context (Y=C/T; H=A/C/T), and also an unexpected prevalence of T>A transversions in a TpTpA context and T>C transitions in TpTpN context (Figure 5.8b). As we used only UV-B for this experiment, we expected more 6-4PPs to be generated (Cadet and Douki 2018). 6-4PPs in *C. elegans* are also repaired slower than in humans (Hartman et al. 1989), and have a higher chance to remain unrepaired until cell division. Hence, a high fraction of T>C mutations can be stemming from the error-prone bypass of 6-4PPs by polymerase η. These changes were further exaggerated in NER deficient mutants (which will be discussed in detail in the next chapter).

In addition, a large fraction of UV-generated mutations across all genotypes were dinucleotide substitutions, in particular, CC>TT/AT (Figure 5.6), similar to the UV-associated DNV spectrum inferred from cancers (Alexandrov et al. 2018). In total, UV-B exposure with the energy of 100 J/m^2 led to approximately 8 mutations per genome.

Moreover, C>T mutations (but not T>C) induced by UV exposure demonstrated a transcriptional strand bias, with 10% more C>T mutations happening on the coding strand compared to the template strand, similar to the bias calculated for the UV-associated signature in melanomas (Alexandrov et al. 2013b). This confirms the high contribution of transcription-coupled NER to the repair of UV-induced CPDs.

Ionising radiation

X-rays and γ irradiation are potent mutagens, which are also frequently used in cancer treatment due to their DNA damaging properties. The ability to direct the exposure using irradiating machines allows focusing the maximal energy of the irradiation precisely at the tumour site while minimising the exposure for healthy tissues. The energy of X- and γ-rays is sufficient to ionise atoms in a cell, inducing oxidative damage, and to introduce single- and double-strand breaks (SSBs, DSBs) in the DNA, which can result in single and dinucleotide substitutions, but also deletions and structural variants (Willers, Dahm-Daphi, and Powell 2004).

Both X-rays and γ-rays are electromagnetic waves composed of high-energy photons, but stems from different sources: γ irradiation is emitted when a radioactive substance undergoes decay whereas X-rays are generated by the device that excites electrons (Baskar et al. 2012). Beta-irradiation, or electron beams, are also used in cancer therapy. Electron beams have higher LET (linear energy transfer) and stronger biological effects than photons, but are also more expensive in use, and are typically applied in treatment of radioresistant tumours (Baskar et al. 2012).

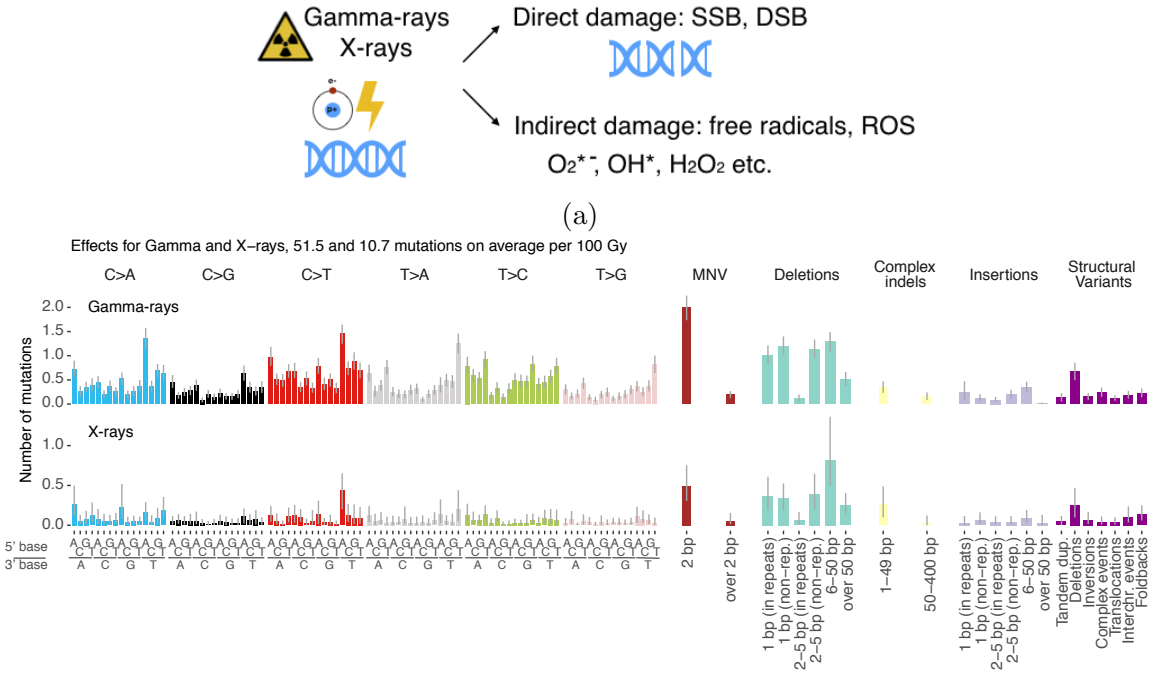


Figure 5.9: Ionising radiation-caused mutagenesis. (a) Damage inflicted by ionising radiation: direct damage on the DNA backbone causing SSBs and DSBs, and indirect damage via production of free radicals and reactive oxygen species (ROS). (b) Experimental mutational signatures of γ and X-ray irradiation in *C. elegans*.

An irradiation dose of 2 Gy can cause up to 3000 lesions in a mammalian cell including 1000 SSBs and only 50 DSBs, the rest being base modifications occurring through the activity of reactive species induced by radiation (Figure 5.9a, Lomax, Folkes, and O'Neill 2013). However, a very small fraction of this damage turns into mutations in healthy cells: exposing mice to 3 Gy of X-rays did not induce elevated numbers of single-nucleotide variants (SNVs), but led to 14 copy number variants (CNVs) compared to just one detected in the untreated control (Russell and Kelly 1982).

In *C. elegans* experiments, Cs-137 was used as a source of γ-rays and, to a lesser degree, β-rays. It induced lethality but not a huge mutational burden in many samples, which indicates the high impact of the DSBs and indirect IR effects – i.e. the induction of oxidative stress – on cell's ability to propagate. On average, Cs-137 exposure led to about 52 mutations per 100 Gy per genome, whereas X-ray yielded 11 mutations per 100 Gy.

The overall distribution of mutations was relatively uniform across all SNV types, without any particular influence of 5' and 3' nucleotides flanking the detected SNVs (Figure 5.9b). Gamma-irradiation but not X-rays also induced a high amount of CC>AA and TC>AA dinucleotide variants (Figure 5.6). These mutations may be arising via guanine intrastrand crosslinks of adjacent bases which can occur at an increased rate upon oxidation stress and irradiation (Cadet et al. 2014). In addition, we observed a high number of deletions, mostly in the spectrum from 1 to 50 bp upon IR exposure but also some longer deletions (Figure 5.9b). This resembles the mutational footprint of BRCA-1/2 deficiency and indicates the NHEJ repair acting on double-strand breaks.

5.2.5 Replication stalling agents – hydroxyurea

Hydroxyurea (Figure 5.10a) is a chemical agent and an inhibitor of the ribonucleotide reductase (RNR) enzyme. As a consequence, no new dNTPs are produced from ribonucleotides. In a dividing cell, it stalls replication forks and ultimately leads to cell cycle arrest when dNTP pool size is below the critical threshold of about 20% necessary for replication (Koç et al. 2004). Stalled replication forks can also collapse into a double-strand break (Figure 5.10b).

As HU does not incur any direct damage to the DNA structure, the mutational signature we observed was unstable, with likely peaks in N[C>A]A and C[C>T]N substitutions which may suggest a particular efficiency of HU in depleting dGTPs. HU only yielded about 1.5 mutations per 1 mM of the agent. Other studies of chemical mutagenesis by HU also did not show a clear mutation pattern under normal conditions (Szikriszt et al. 2016).

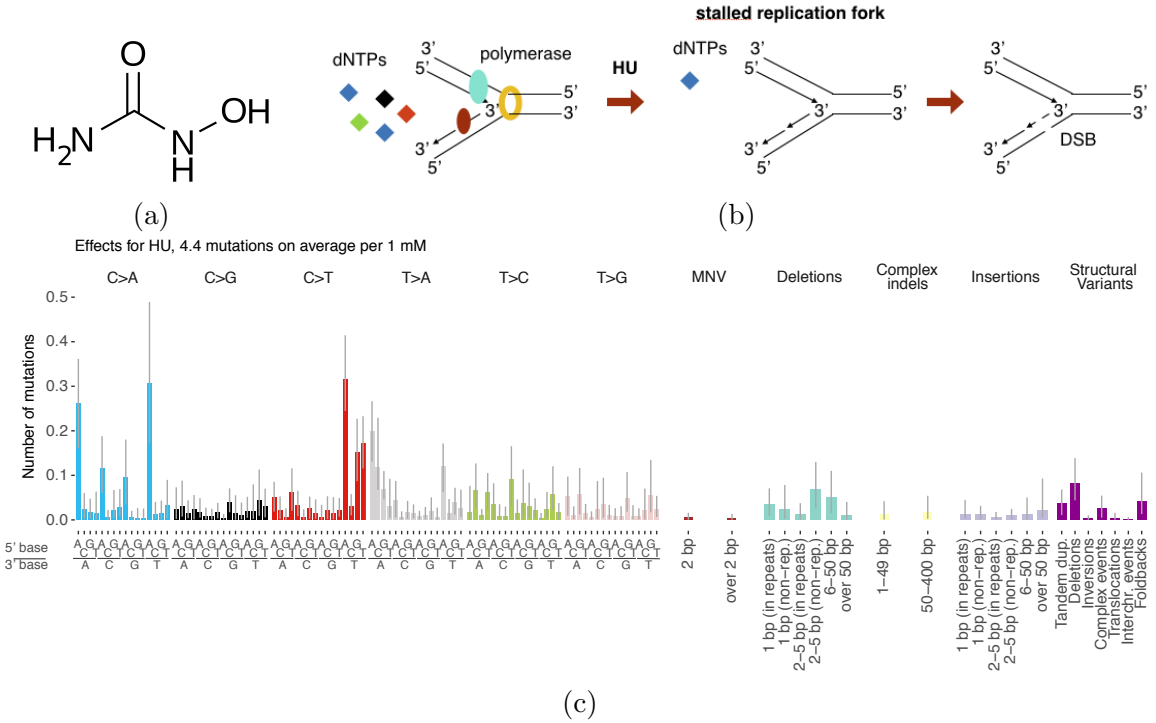


Figure 5.10: Hydroxyurea mutagenesis. (a) Structure of HU molecule. (b) Mechanism of HU-induced depletion of dNTP and replication stalling. Adapted from (c) Experimental mutational signatures of hydroxyurea exposure in *C. elegans*.

5.3 Cross-species comparison

Using the adjustment procedure described in Chapter 2, we assessed the comparability of experimental mutational signatures of different genotoxins in *C. elegans* and human cells (as per Kucab et al. 2019) as well as the similarity between experimental signatures and computational mutational signatures extracted from cancer genomes (Alexandrov et al. 2018).

The signature of EMS in worms was strikingly similar (0.90) to the cancer-derived computational signature SBS11 associated with temozolomide treatment. However, it turned out to be different from the temozolomide signature in iPS cell lines (Figure 5.11a). This may be due to cell type-specific metabolic activation of the mutagen: mutational spectrum observed upon EMS exposure in *Salmonella typhimurium* was nearly identical to the one we saw in *C. elegans* (Matsumura et al. 2018).

UV-B exposure in worms showed a mutation spectrum dominated by C>T, which was similar to that in cell line experiments and cancer, albeit with an additional fraction of T>C mutations (Figure 5.11b). As discussed above, the discrepancy can be coming from the difference in UV exposure sources: we used a UV-B source, whereas Kucab et al. 2019 used a mixture of 90% UV-A and 10% UV-B similar to actual UV spectrum contained in

the sunlight.

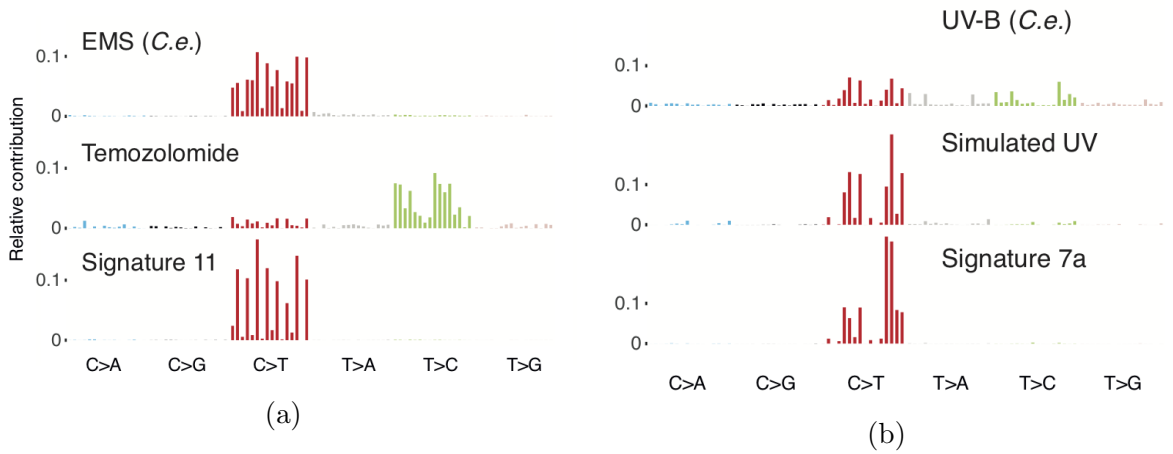


Figure 5.11: Experimental signatures of (a) EMS and (b) UV exposures in *C. elegans* (top) and human iPS cells (middle), and cancer signatures (a) SBS11 and (b) SBS7a (bottom).

The signature of aflatoxin-B1 was similar to the computationally extracted signature SBS24 (similarity 0.8) but different compared to the experimental one (0.62) (Figure 5.12a). Similarly, aristolochic acid signatures were consistent between both experimental systems and cancer, with cosine similarities of 0.91 and 0.90, respectively (Figure 5.12b).

As there was no experimental data available for gamma-irradiation in human cells, we considered an averaged spectrum of 12 radiotherapy-associated secondary malignancies from Behjati et al. 2016, which showed a high similarity to the *C. elegans* Cs-137 induced base substitution signature. The spectrum of indels associated with this exposure also agreed with the prevalence of small indels found in Behjati et al. 2016. Compared to computationally derived cancer signatures, IR exposure spectrum was most similar to SBS40 (Figure 5.12d), which is found across all cancer types (Alexandrov et al. 2018). It suggests that the SBS40, as well as the IR spectra, are generated by error-prone repair of DSBs. Interestingly, the spectrum of IR-induced mutations observed in *C. elegans* was different from an IR signature found in a pooled analysis of human and mouse radiation-induced malignancies (Davidson et al. 2017), which was mostly characterised by C>A,T mutations.

Signature of cisplatin treatment in *C. elegans* consisted almost exclusively of C>A mutations being different from the one identified in human iPS cells, which was shifted towards C>T contribution. Mutational spectra observed upon cisplatin treatment in chicken fibroblasts, however, had a similar preference for C>A mutations (Szikriszt et al. 2016). Multiple studies of cisplatin-treated cancers suggested that cisplatin treatment yields both C>A and C>T mutations (Boot et al. 2018, Liu et al. 2017a), which also

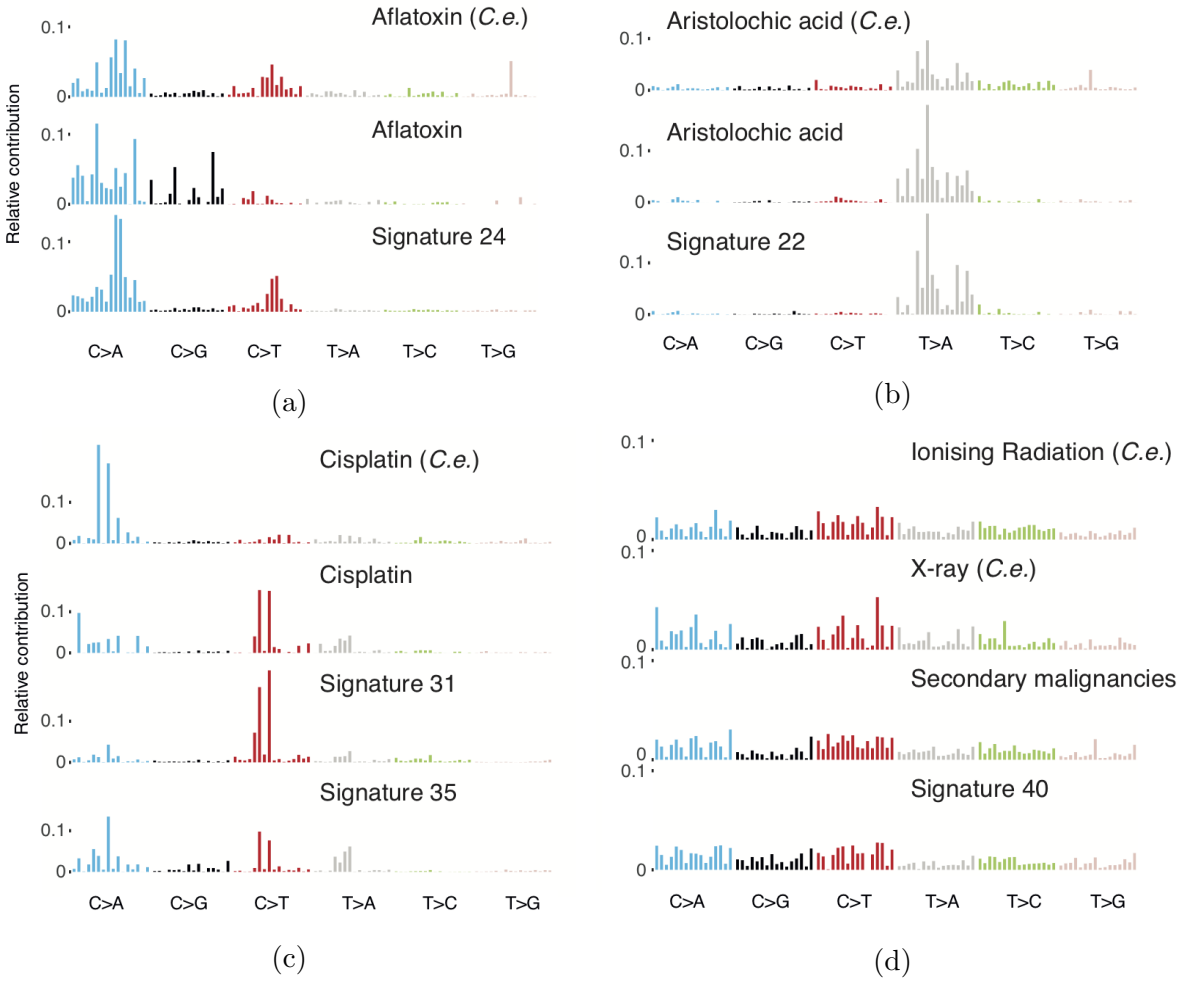


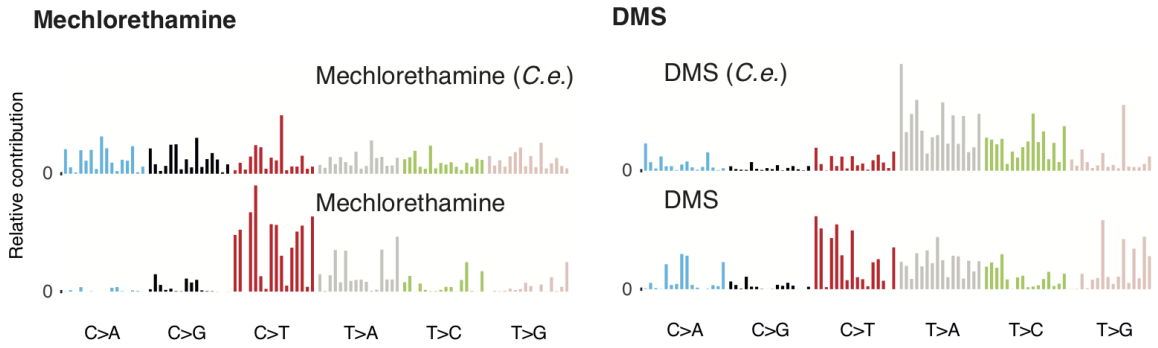
Figure 5.12: Experimental signatures of (a) aflatoxin-B1, (b) aristolochic acid, (c) cisplatin and (d) γ-rays and X-rays exposures in *C. elegans* (top) and (a)-(c) human iPS cells or (d) average mutational spectrum of radiation-associated secondary malignancies from Behjati et al. 2016 (middle), and corresponding COSMIC cancer signature (a) SBS24, (b) SBS22, (c) SBS31 and SBS35 and (d) SBS40 (bottom).

confirms the link between signatures SBS31/35 and platinum-based agent treatment.

Exposures to mechlorethamine and DMS exhibited different mutational spectra in the two model systems (Figure 5.13).

No counterparts from human data were found for HU, Mitomycin C, and MMS.

Thus, we observed *C. elegans* signatures matching the ones in derived from cancer samples for half of the genotoxins tested. This similarity reflects the fact that the majority of DNA repair pathways are highly conserved among eukaryotes, but also that DNA repair capacity and genotoxin metabolism may differ moderately between nematodes, human cell lines, and cancer cells. Human cell lines were, in fact, no better model system for cancer: some of the experimental signatures derived from human cells, such as signatures of aflatoxin (Figure 5.12a) and temozolomide (Figure 5.11a), were further away from cancer



(a) (b)

Figure 5.13: Experimental signatures of (a) mechlorethamine and (b) DMS exposures in *C. elegans* (top) and human iPS cells (bottom).

signatures than *C. elegans* ones.

At a mean cosine similarity of 0.63 (range 0.20-0.84), these experimental signatures generally displayed a good, although not perfect level of similarity with their human experimental counterparts and also with computationally derived cancer signatures with the same suspected origins (Figure 5.14).

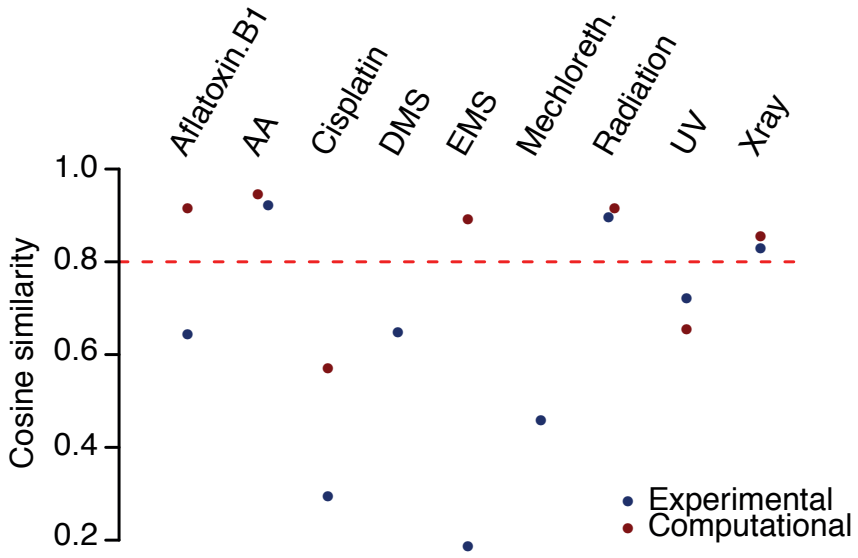


Figure 5.14: Similarities between mutational signatures of genotoxins in human and in *C. elegans*.

5.4 Discussion

In this chapter, I introduced the genotoxins used in the screen and demonstrated the signatures generated by these substances upon exposure in wild-type *C. elegans*. Experimental signatures of mutagens with different mechanisms of mutagenicity were described on a genome-wide scale, demonstrating how different mechanisms of interaction with DNA generate different spectra of mutations.

Many of the genotoxins used in mutagenesis experiments can yield a high amount of lesions, yet result in a small number of mutations if those lesions are not mutagenic or can be efficiently repaired. Alkylating agents EMS and MMS produce more alkylated bases than mutations. The distribution of these mutations also does not necessarily reflect the distribution of damage as some DNA damage can be repaired faster or better than other, such as UV-induced pyrimidine dimers at different dinucleotide contexts. Hence, these differences can indicate differential mutagenic potency, such as TLS-mediated mispairing during replication, or differential repair, as in the case of mechlorethamine.

In addition, I compared the experimental signatures in *C. elegans* to those derived from human iPS cell line and cancers. Spectra of aristolochic acid and ionising radiation were well conserved, whereas other agents such as cisplatin and mechlorethamine showed diverging distributions of mutations. Interestingly, signatures associated with aflatoxin and temozolomide exposures in cancers were more similar to the spectra obtained in the corresponding *C. elegans* experiments than human cell lines. This overview confirmed that *C. elegans* is a good model to study the mechanisms of mutagenesis due to its simplicity, but the limited capacity of cross-species comparison revealed several limitations of this model system.

As discussed before, one of the drawbacks of *C. elegans* is its low tolerance to mutations due to the high proportion of coding genome and, consequently, high probability of deleterious mutations. It limited the maximal dose which could be applied to obtain mutations, which was compensated by multiple replicates which were pulled together to infer the experimental mutational signatures.

A general challenge to the studies of mutagenesis in model systems are the differences in the metabolism of different genotoxins. Multiple alkylating agents showed a different mutational spectrum in human cells compared to *C. elegans* probably indicating different reactive compounds and corresponding damage mechanisms, but the agreement between the iPSCs and actual cancers was also limited (Kucab et al. 2019). The discrepancies in mutational spectra across model systems can provide information about the complexity of metabolic processing of genotoxic substances.

Divergent mutational spectra arising in response to the same genotoxin can also in-

dicate different DNA repair pathways acting upon the inflicted DNA damage. Among these, a wider range of translesion synthesis polymerases in humans compared to nematodes could lead to a different error-prone repair of the same lesions: for instance, polymerase ι can perform mutagenic bypass of UV-lesions in humans but is absent in *C. elegans* (Knobel and Marti 2011). Hence, it is important to remember that a genotoxin only suggests a set of DNA lesions, but the observed mutational spectrum is ultimately shaped by the activity of DNA repair and tolerance mechanisms.

A more systematic assessment of the rates and genetic determinants of DNA lesions in relation to the resulting mutation spectra could shed more light on the mutagenicity and repair efficiency of different DNA damage. To address the latter aspect, namely the efficiency of DNA repair and its potential to alter the mutagenic properties of genotoxins, we analysed the mutational spectra generated by the same agents across a range of DNA repair-deficient backgrounds. In the next chapter, I will study the relationship between DNA damage and repair in more detail, present a model for quantification of the mutagenic contribution of genotypes and genotoxins, and demonstrate the frequency and magnitude of interaction effects across a range of combinations between DNA damage deficiencies and DNA damage types.

Chapter 6

Interactions between DNA damage and repair

6.1 Introduction

In the previous chapters, I described the genome-wide mutational signatures of DNA repair deficiencies and genotoxin exposures in *C. elegans*. Mutations observed upon sequencing typically arise in a two-step manner: it starts with damage to the DNA, and it is the repair or replication over a lesion which leads to mutation. Given this double-sided nature of mutation acquisition, it seems reasonable to assume that changes in the DNA repair component availability may affect the mutational signatures of exogenous or endogenous mutagenic agents.

In this chapter, I will quantify and describe the contributions of different factors and their interactions to the mutational spectra of samples with combined DNA repair deficiency and mutagen exposure. I will also show that the interplay between DNA repair and damage is common and can alter the signature of the mutagen due to a switch of DNA repair pathway that acts on the damage, and present a summary of interaction effects along with the most striking examples. This chapter will further stress the fact that the strongest mutational signals – which would be the first ones to be picked up by conventional unsupervised factor analysis methods – usually result from non-linear interactions of different factors, and should not be considered as caused by a single independent factor.

Contributions

The findings described in this chapter were submitted for publication as a part of the following manuscript:

Volkova, N.V., Meier, B., González-Huici, V., Bertolini, S., Gonzalez, S., Abascal,

F., Martincorena, I., Campbell, P.J., Gartner, A. and Gerstung, M. (2019). Mutational signatures are jointly shaped by DNA damage and repair. *bioRxiv*, 686295.

This chapter represents a reformulation of the first part of the manuscript, which focuses on the DNA damage-repair interactions in the *C. elegans* experimental data. The manuscript was restructured to highlight the methodological innovation as follows: section 6.2 provides a novel introduction presenting the interaction concept, and section 6.3 is almost identical to the one in the manuscript and describes the most striking examples and a summary of the frequency and magnitude of the damage-repair interactions across the mutagenesis screen.

This work was conducted in collaboration with Bettina Meier and colleagues from Anton Gartner's research group at the University of Dundee, and Inigo Martincorena's and Peter Campbell's groups at the Sanger Institute. All of the analyses described below were performed by me.

6.2 Interplay between DNA repair and DNA damage

The genetic material of a cell is constantly being attacked by various types of damaging processes. Replicative errors induce mismatches or contraction/expansion of repetitive regions, which result in mutations in one of the daughter cells. Replication-transcription collapses can lead to replication fork collapse and rearrangements of DNA. Reactive species within the cell react with different nucleotides and DNA backbone, inducing base modifications and strand breaks. The presence of additional mutagenic processes, such as exposures to environmental genotoxins, further amplifies the number of lesions that cell's DNA repair machinery has to deal with.

Many types of DNA lesions can be processed by multiple DNA repair pathways. The redundancy of DNA repair ensures persistence of replication and survival of the organism. However, different DNA repair pathways may have different accuracy and efficacy when repairing the same damage, which means that the mutational footprint left by the same damaging agents will not be the same when different DNA repair components act upon it.

For example, exposure to benzo-[a]-pyrene, one of the main components of tobacco smoke (Phillips 2002), leads to the formation of bulky adducts on guanines (Li et al. 2017a). Typically, these adducts would be excised and correctly replaced with a G by nucleotide excision repair (Hess et al. 1997). However, if the lesion persists until the replication, it can be replicated over with translesion synthesis polymerases (Rechkoblit et al. 2002), which may result in a G>A mutation in one of the daughter cells. Alternatively, if left unrepaired, BaP-adducts can cause double-strand breaks by altering the spatial

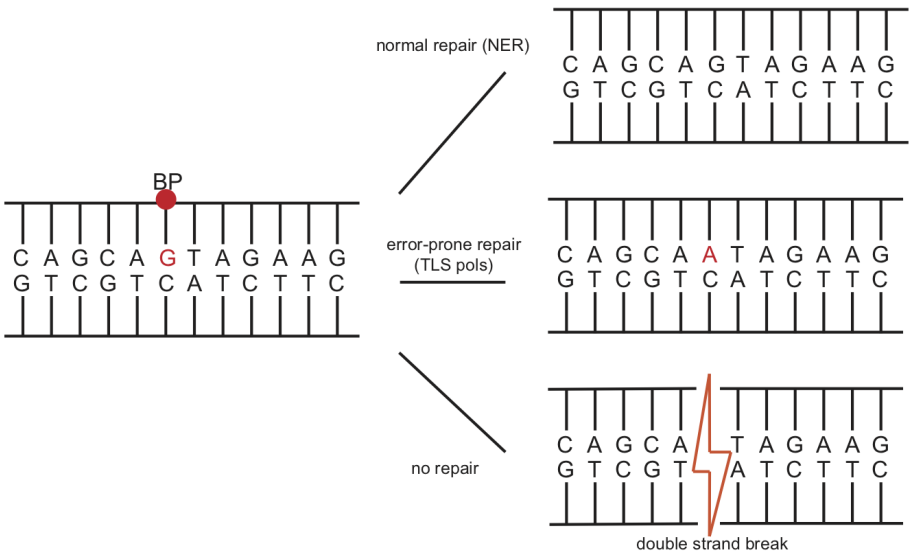


Figure 6.1: The concept of damage-repair interaction. The precise type of mutation observed depends on the repair available. BP - Benzo-[a]-pyrene adduct.

structure of the DNA around the adduct (Tung et al. 2014). Both cases lead to different variants, caused by the same genotoxin. Hence, we suggest that the DNA repair status can alter the mutational signature of a mutagen.

This implies that there are at least two unknowns that contribute to a mutational spectrum. This dichotomy holds not only for exogenous mutagens but also for any cellular process which can introduce DNA damage. The effects of such interactions are exemplified by cancers with combined MMR deficiency and defects in exonuclease domain of replicative polymerase ϵ, which display a profoundly different spectrum compared to samples with the same POLE deficiency but intact MMR (Haradhvala et al. 2018).

Measuring such interaction requires knowledge of the nature of DNA damage and the cell's repair capacity. We, therefore, conducted a large combinatorial mutagenesis screen, which allowed us to deconvolve the contributions of genotoxic processes and DNA repair machinery to the samples where a knockout and an exposure were combined. Experiments combining genotoxin exposure and DNA repair deficiency show signs of altered mutagenesis, signified by either higher or lower rates of mutations as well as altered mutation spectra. These interactions highlight how DNA lesions arising from the same genotoxin are mended by a number of DNA repair pathways, often specific towards a particular type of DNA damage and therefore changing mutation spectra usually in subtle but sometimes also dramatic ways.

6.3 Quantifying interaction effects in a controlled experimental model

The controlled nature of the mutagenesis screen allowed us to take a step further towards dissecting the contributions of DNA repair and DNA damage to the mutational spectra of *C. elegans* samples.

The experimental setup provided us with the full information about the absolute dosage of each genotoxin and the duration of the exposure to cell-intrinsic mutagenic processes for each sample. With this data, it was possible to go beyond the recognition of recurrent patterns and quantify the actual amount of change introduced by each interaction. Hence, our analysis could capture both amplifications and reductions of existing signatures as well as transformations of genotoxin-induced signatures. This is unlike the conventional signature analysis, which can only model additive effects because additional factors are assumed to be solely able to add mutations; however, the data from our experiments indicated that mutagenesis could also be reduced.

In order to fulfil this task while maintaining interpretability of all model components, we developed a hierarchical Bayesian model which adapts a range of the possible changes and noise sources in the data, discussed in Chapter 2.

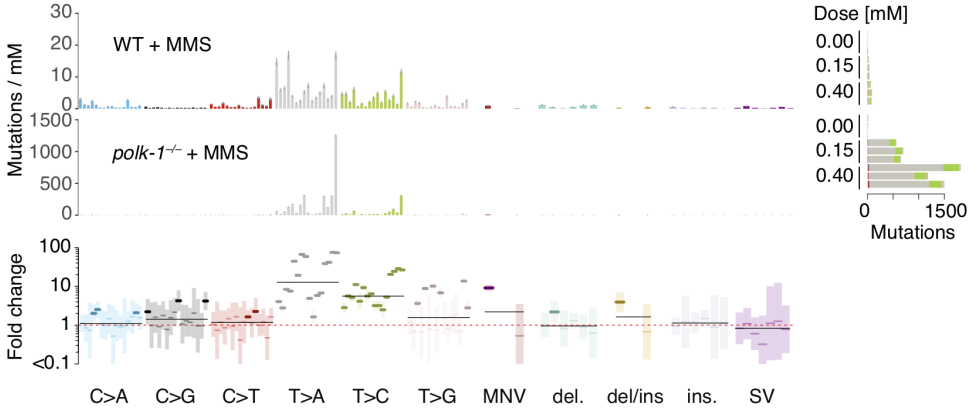
6.4 Alteration of mutagen profiles in *C. elegans* experiments

Upon quantifying the interaction effects, we confirmed a number of known cases of the interplay between DNA damage and repair, for which we now offered a genome-wide spectrum estimation, as well as identified previously understudied combinations.

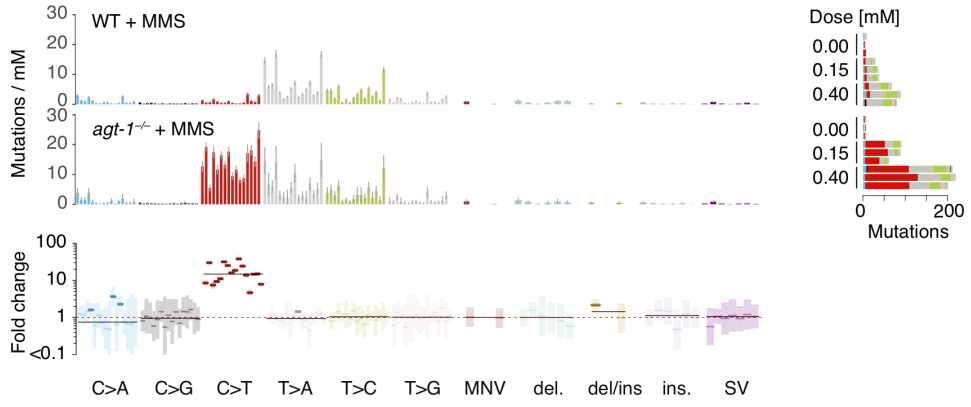
6.4.1 Alkylating agents and corresponding repair enzymes

Among all of the interaction experiments, the highest number of mutations was observed for the knockout of TLS polymerase *polk-1* under exposure to alkylating agent MMS. Deficiency of polymerase κ increases the total mutation rate 17x to approximately 3,800 mutations/mM/generation (Figure 6.2a). This increase also coincides with a distinct change in the mutational spectrum with a prominent peak of T>A transversions in a TpTpT context (Figure 6.2a, central panel).

Quantifying these changes relative to the wild-type mutation spectrum and accounting for possible genotoxin-independent effects of polymerase κ deficiency revealed that the rate of T>M mutations is approximately 10-100x higher, especially in TpTpN and CpTpN



(a)



(b)

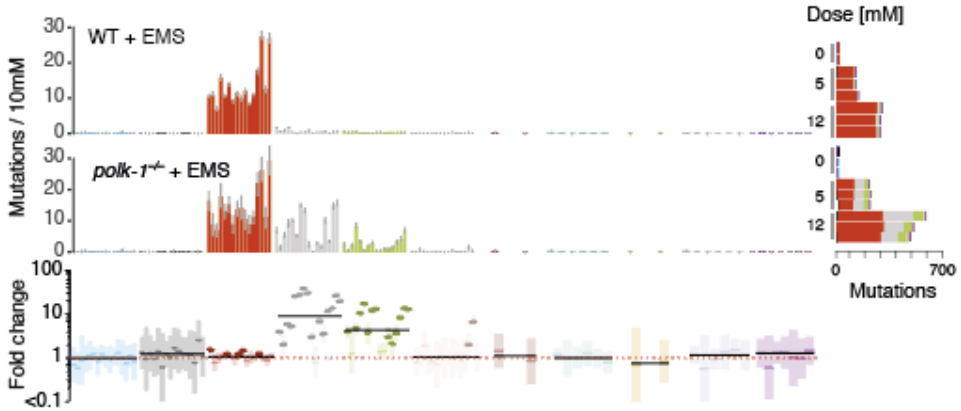
Figure 6.2: Mutations introduced per unit of MMS in wild-type (top panel) and in (a) *polk-1* or (b) *agt-1* deficient mutants (central panel), along with the fold-change per mutation type (bottom) and the total numbers of mutations in response to different doses (right panel). Corresponds to Figure 3A/B in Volkova et al. 2019.

contexts (Figure 6.2a, bottom). These figures indicate that 90-99% of MMS-induced mutations are avoided by Pol κ driven TLS. T>M mutagenesis is likely to be caused by N3-methyladenosine, which is considered to represent the largest proportion of mutagenic adducts, stalls replicative polymerases, and is mended by several TLS polymerases (Yoon et al. 2017). Our data indicate that TLS in the *polk-1* mutant, contrary to the wild-type, N3-meA bypass had to be achieved by other, error-prone TLS polymerases, resulting in largely increased T>M mutagenesis.

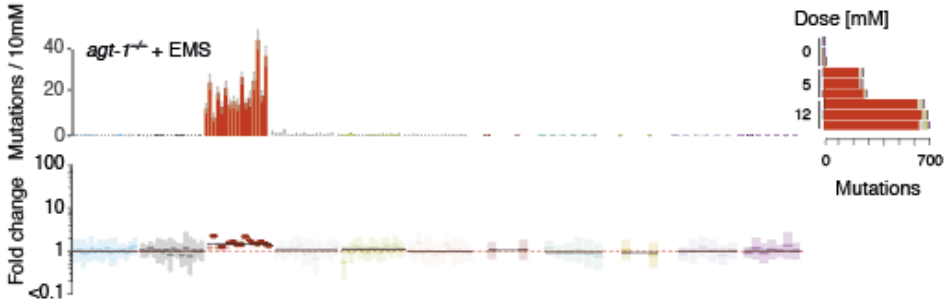
A further mutagenic DNA modification induced by MMS is O6-methylguanine, which leads to G:T mispairing resulting in C>T transitions. In wild-type experiments, MMS induced less than 10 C>T mutations/mM. Combining MMS exposure with alkyl-transferase *agt-1* deficiency, however, increases C>T mutations by a factor of 12, while leaving the rate of T>M mutations unchanged (Figure 6.2b). This demonstrates that AGT-1 specif-

ically reverses most O6-methyl-guanine adducts, thus acting as the functional *C.elegans* ortholog of the human O6-methylguanine DNA methyltransferase MGMT.

A similar change of the mutation spectrum was observed for DNA ethylation by EMS. Upon exposure in *polk-1* deficient samples, EMS signature demonstrated a 10x increase in T>M mutations (Figure 6.3a), whereas in *agt-1* mutants it only showed a 1.5-fold increase in C>T mutations (Figure 6.3b). These changes in the signatures suggest that the adenine alkylation damage inflicted by MMS and EMS is normally bypassed via POLK-1 mediated TLS, and that AGT-1 efficiently repairs guanine methylation, but is less efficient in repairing guanine ethylation, which is consistent with the reports in *E. coli* (Taira et al. 2013).



(a)



(b)

Figure 6.3: (a) Mutations introduced per unit of EMS in (a) wild-type (top panel) and in *polk-1* deficient mutants (central panel), or in (b) *agt-1* deficient mutants (top panel), along with the fold-change per mutation type (bottom) and total numbers of mutation in response to different doses of mutagen (right panel). Corresponds to Supplementary Figure 4 in Volkova et al. 2019.

6.4.2 Translesion synthesis deficiency decreases the number of observed mutations

Knockouts of two other translesion synthesis polymerases, POLH-1 and REV-3, led to a severe inability of the cell to replicate over multiple sorts of lesions. As demonstrated in Chapter 3, the mutation accumulation patterns for these knockouts in the absence of exogenous genotoxicity were characterised by deletions in the range of 50-400 bp, thought to arise via the formation of DNA double-strand breaks and *polq-1* mediated repair.

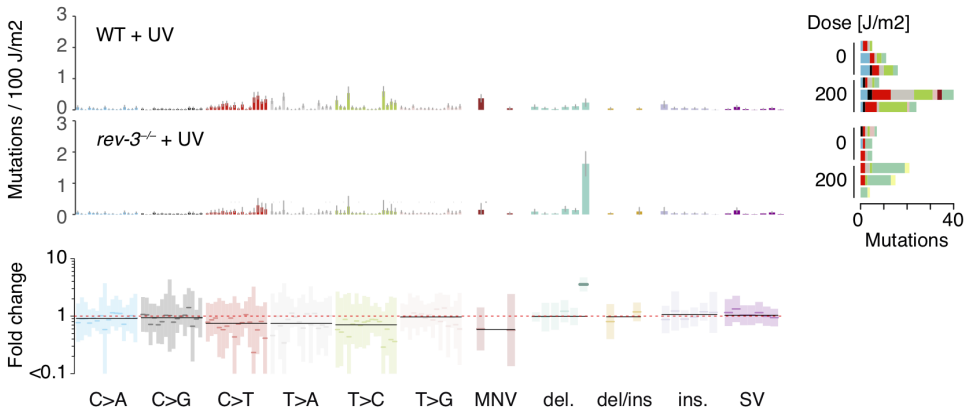
Error-prone TLS is a key mechanism to overcome UV-induced DNA damage such as cyclobutane pyrimidine dimers, which stall replication polymerases. Exposure of TLS polymerase η and ζ mutants, the latter one in particular, led to a distinctive reduction in the amount of observed substitution: upon UV exposure, *rev-3* mutant showed a 1.5-fold reduction in substitutions and 4-fold increase in deletions $>$ 50 bp (Figure 6.4a).

MMS exposure in *rev-3* mutants caused severe lethality, making the mutants unable to bear the doses higher than the lowest one used in the wild-type experiments (50 μM of MMS), which led to a very small and variable number of substitutions. Projecting the values of mutation burden observed at low mutagen concentrations to those used for the wild-type samples, we expect a 20% reduction in the number of substitutions and a 5-fold increase in the number of indels (Figure 6.4b). This supports the concept of TLS bypass protecting the genome from more severe mutation - such as deletions - at the cost of increased error rate and consequently elevated base substitution rate (Yoon et al. 2019).

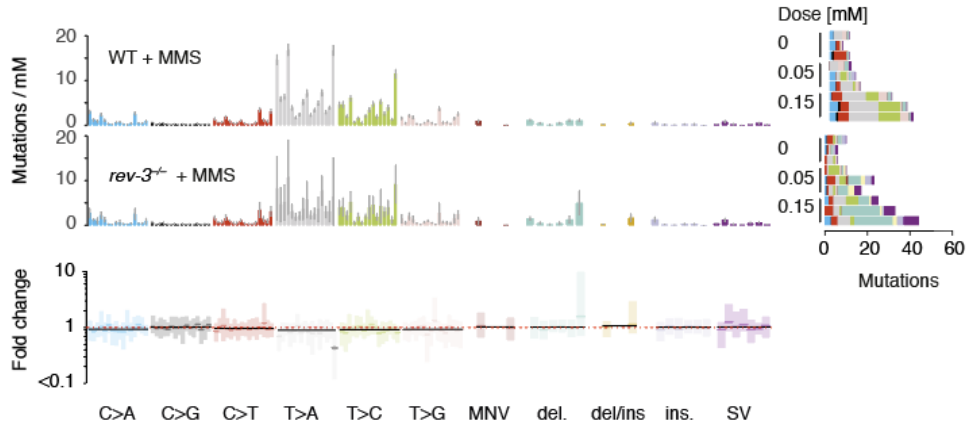
6.4.3 Nucleotide excision repair deficiency exacerbates the effects of mutagens

In contrast to the above examples where repair deficiency leads to both increase in mutagenesis but also to alteration of mutational spectra, the knockouts of the NER genes *xpf-1* and *xpc-1* increased the rate of UV-B induced mutagenesis by factors of 10 and 35, respectively. The fold-change appears to be relatively uniform across the entire mutation spectrum, indicating that NER is involved in repairing the large majority of UV-B damage of different types, including both single- and multinucleotide lesions (Figure 6.5a).

Interestingly, *xpf-1* and *xpc-1* knockouts also uniformly increased the mutation burden due to alkylation by a factor of 2, indicating that alkylation damage is eventually repaired by NER (Appendix B). Similarly, both *xpf-1* and *xpc-1* knockouts also showed a two-fold increase in mutations after aristolochic acid, demonstrating that NER repairs both small and bulky adducts (Figure 6.5). Against previous reports on the lack of recognition of the aristolactam adducts by global genome NER (Sidorenko et al. 2012), *xpc-1* mutants in our screen showed an increase in mutations upon aristolochic acid exposure, especially



(a)



(b)

Figure 6.4: Mutations introduced per unit of (a) UV irradiation and (b) MMS in wild-type (top panel) and in *rev-3* deficient mutants (central panel), along with the fold-change per mutation type (bottom) and the total numbers of mutations in response to different doses (right panel). Corresponds to Figure 3C and Supplementary Figure 4 in Volkova et al. 2019.

deletions in the range of 50-400 bps (Figure B), similar to that in *xpf-1* mutants deficient in both NER modalities. This may be a consequence of the experimental setup, in which *C. elegans* has to survive the embryonic development stage, when GG-NER is more active than in adult stages (Lans and Vermeulen 2011).

A strong change in the mutational spectrum was also observed upon irradiation of *xpa-1* deficient mutants with γ-rays (Figure 6.5c). We observed a 1.5-fold increase in C>T,A base changes and indels, as well as a 3-fold increase in DNVs. XPA is a NER factor involved in damage recognition for both TC-NER and GG-NER, which also serves as a scaffold to assemble the excision complex. Similar effects were observed in other NER mutants, but to a lesser degree (Appendix B). It is likely that these proteins enhance the damage detection and repair capacity of DSB repair (Zhang, Rohde, and Wu 2009),

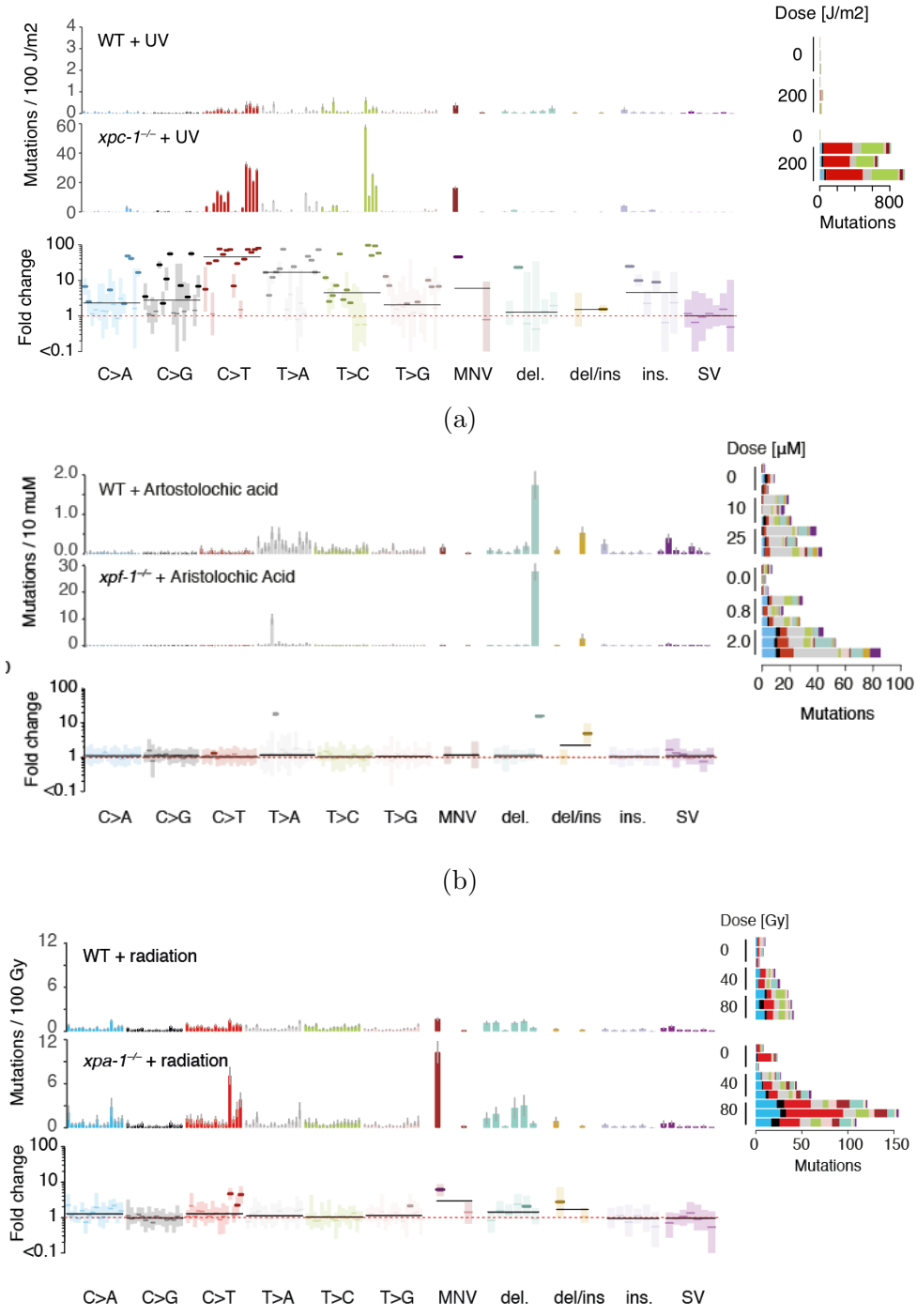


Figure 6.5: Mutations introduced per unit of (a) UV-irradiation, (b) aristolochic acid exposure or (c) γ-irradiation in wild-type (top panel) and in (a) *xpc-1*, (b) *xpf-1* and (c) *xpa-1* deficient mutants (central panel), along with the fold-change per mutation type (bottom) and total numbers of mutations in response to different doses of mutagen (right panel). Corresponds to Figure 3D and Supplementary Figure 4 in Volkova et al. 2019.

and an absence of XPA can increase the amount of unrepaired SSBs and intrastrand crosslinks, which result in DNVs, and mutagenic DSB repair generating base substitutions and deletions.

6.5 Widespread and potent damage-repair interactions in *C. elegans* screen

These examples of genotoxin-repair interactions are not uncommon: in total, 72/196 (37%, at FDR of 10%) of combinations displayed an interaction between DNA repair status and genotoxin treatment, involving 9 out of 11 genotoxins which had interaction experiments (Figure 6.6). Conversely, more than a half of combinations produced mutation spectra which could be fully described as the superposition of the wild-type genotoxin signature and that of the DNA repair deficiency background, indicating that the two processes acted independently. Usually, genotoxin-repair interactions increase the numbers of mutations obtained for a given dose of mutagen, leaving the mutational spectrum mostly unchanged; others have a profound impact on mutational spectra (Figure 6.7).

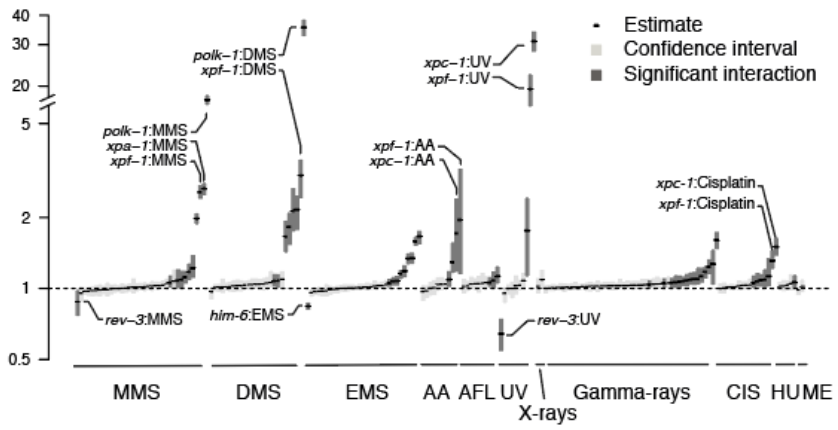


Figure 6.6: Fold changes in base substitution rates between the effects of mutagen exposure in the wild-type and under different DNA repair-deficient conditions. Each bar marks a 95% confidence interval for the fold-change of the base substitution rate. Dark bars denote the combinations with a fold-change significantly different from 1. Corresponds to Figure 4A in Volkova et al. 2019.

The emergence of distinct mutational spectra depending on DNA repair status reflects how multiple repair enzymes and pathways contribute to mending damaged DNA as in the case of DNA alkylation. If DNA repair is predominantly achieved by one pathway such as by NER for UVB damage and bulky DNA adducts, the number of mutations is increased, but the signature tends to remain unchanged.

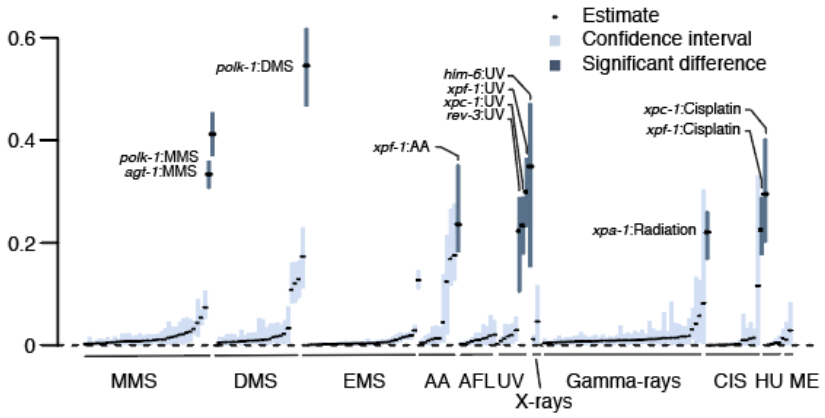


Figure 6.7: Cosine distances between the mutational signatures of mutagens in the wild-type and under different DNA repair-deficient conditions. Each bar marks a 95% confidence interval for the distance between the signatures. Dark bars denote the combinations with a mean distance higher than 0.2 (the threshold was chosen based on the simulations, Appendix **??**). Corresponds to Figure 4B in Volkova et al. 2019.

To illustrate the overall magnitude of these interaction effects, we estimate that of the 141,004 mutations we observed upon treatment with genotoxins in DNA repair-deficient strains 23% of mutations were attributed to the endogenous mutagenicity of DNA repair deficiency genotypes independent of genotoxic exposure, and 62% of mutations would be attributed to genotoxic exposures independent of the genetic background. Of these, 2% were not observed due to negative interactions of genotoxins and DNA repair deficiency, such as UVB exposure of TLS knockout strains, and 17% of mutations were added because of the positive interactions, which increased mutagenicity (Figure 6.8).

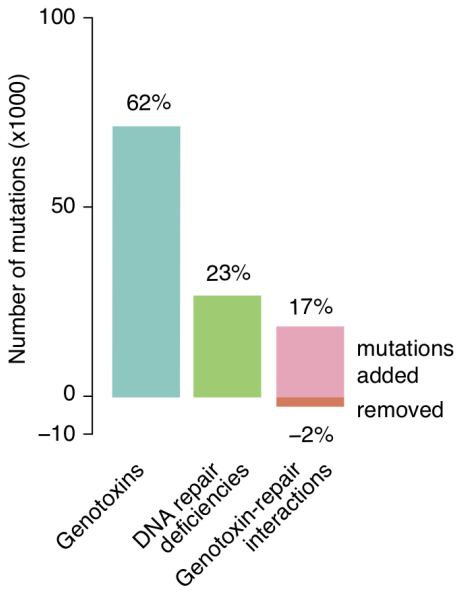


Figure 6.8: Contributions of different factors across all interaction experiments. Corresponds to Figure 4C in Volkova et al. 2019.

6.6 Discussion

In this chapter, I presented a way of quantifying the degree of interaction between genetic and mutagenic factors in controlled experiments on *C. elegans*, and performed a comparison of effects across 196 combinations of repair components and damaging agents which showed the diversity and abundance of significant

interaction effects which may be both positive and negative. I also demonstrated how these interactions might increase or decrease the total burden, and alter the signature of the mutagenic agent as well as intensify the signature of genetic factors, which in total was observed in 37% of experiments.

The mode of interaction detected for a mutagen in a DNA repair-deficient environment indicates the relationship between the lesions introduced by the damaging agent and the repair pathway. Observing a change of the spectrum, as in the case of alkylating agents and *polk-1* and *agt-1* knockouts suggests that divergent repair pathways are involved in repairing different lesions introduced by the same genotoxin. When either repair component is knocked out, a distinct part of the mutation spectrum changes, indicating those lesions which would otherwise be repaired by the given repair pathway. Conversely, a change in exposure only indicates that the inactivated DNA repair component is either contributing to repair of all the lesions introduced by a damaging agent, as in case of UV exposure and NER deficient mutants or participates in the damage response pathway.

Overall, this analysis shows that the signatures of many mutagenic processes are not necessarily constant, and may vary in different ways. As multiple and only partially overlapping repair pathways are mending genotoxic lesions, the resulting mutation spectrum may change depending on the activity of each component in a given cell. It opens the way for considering DNA repair status as a factor in signature decomposition as it may alter the appearance of latent components or introduce additional ones.

Studying the effects of DNA repair deficiencies on the mutagenic exposures also provides insights into the mechanisms of repair and specificity of the damage. We observed a shift in context preference of mutational signature of MMS exposure upon *polk-1* deficiency, leading to a peak in T[T>A]T mutations. It raises the question of whether it is due to MMS preferentially introducing damage in this context, or due to POLK-1 being able to replicate over a lesion in this context much more accurately than other TLS polymerases?

The screen above only considers mutations, i.e. results of repair-damage interplay; combined with the analysis of unrepaired damage, which may be detected with direct damage detection methods based on fluorescence or spectrometry (Sykora et al. 2018, Figueroa-González and Pérez-Plasencia 2017), it could provide a comprehensive and quantitative picture of mechanisms of damage and repair of the genomic DNA. Techniques sensitive to any kind of damage, however, can only provide the intensity of damage but not its precise genomic location. More specific methods designed to detect a particular type of lesions are being developed including XR-seq for UV-induced CPD detection (Hu et al. 2015) and AP-seq for oxidative damage detection (Poetsch, Boulton, and Luscombe 2018), as well as computational methods for damaged base detection in Nanopore

sequencing (Liu et al. 2019, Georgieva et al. 2019).

The pervasive existence of mutagen-repair interactions raised the question to which extent similar phenomena can be observed in human cancers. In the next chapter, I will explore the range of repair-damage interaction in human cancers by modifying the signature extraction model and incorporating additional information about DNA repair status of the samples.

Chapter 7

Interplay between DNA damage and repair in cancer

7.1 Introduction

In the previous chapter, I have introduced the idea of mutational signatures being a product of interplay between DNA damage and DNA repair and described the most striking examples of such interactions observed in the *C. elegans* mutagenesis screen.

Translating these findings to cancer requires a way of confidently identifying the presence and degree of DNA repair or DNA damage response impairment based on the genetic variants or epigenetic silencing. Establishing whether a repair pathway is genetically impaired requires estimating the pathogenic consequences of point mutations, as well as there allelic status. For example, for POLE exonuclease domain variants or MMR defects, a heterozygous damaging variant would be sufficient. However, other DNA repair deficiencies would only be exhibited upon biallelic loss of function, conferred either by two independent mutations, or by a loss of heterozygosity after heterozygous mutation, or via a combination of a mutation and gene silencing (Lahtz and Pfeifer 2011).

Secondly, it would require knowledge of the origins and degree of all the genotoxic exposures in a given sample, which are generally unknown. Certain factors such as exposure to UV light, smoking or consumption of genotoxins contained in food may be determined and quantified using clinical data and patient reports, but the actual exposure in each tissue may vary substantially in response to many biological and environmental covariates such as ethnicity or lifestyle choices (Ward et al. 2004). For example, the incidence of skin cancers resulting from UV-induced damage the same amount of exposure would be dramatically different in people of Caucasian or African origin due to different amounts of skin pigmentation (Brenner and Hearing 2008), similarly for people residing in northern

Europe or Australia (Rivas et al. 2011), or those who use sunscreen or not (Green et al. 2011). Similarly, representatives of African Americans and Native Hawaiian ethnicity in the USA have a higher smoking-associated lung cancer risk than the white population adjusted for the number of cigarettes per day (Haiman et al. 2006).

Nevertheless, knowing about the variability of mutational signatures presents the means for a more informed signature analysis in individual samples. Thus, in this chapter, I will explore the frequency of DNA repair deficiencies across human cancers and the overall contribution of defects in DNA repair-related genes to cancer development. I will also present a statistical model for the simultaneous extraction of mutational signatures and estimation of the effects of DNA repair inactivation on the appearance of these signatures, which will allow for analysing damage-repair interactions in tumours with strong mutagenic components.

Contributions

The findings described in this chapter were submitted for publication as a part of the following manuscript:

Volkova, N.V., Meier, B., González-Huici, V., Bertolini, S., Gonzalez, S., Abascal, F., Martincorena, I., Campbell, P.J., Gartner, A. and Gerstung, M. (2019). Mutational signatures are jointly shaped by DNA damage and repair. *bioRxiv*, 686295.

This chapter represents a reformulation of the second part of the manuscript, which focuses on the DNA damage-repair interactions in human cancer. The text of the manuscript was restructured to present a full overview of DNA repair deficiencies in cancer. Section 7.2 provides an overview of frequencies and consequences of DNA repair deficiencies on the mutational burden and spectra of tumours, section 7.3 summarises the method and presents the results for interactions tested.

This work was conducted in collaboration with Bettina Meier and colleagues from Anton Gartner's research group at the University of Dundee, and Peter Campbell at the Sanger Institute, who conceived the study. Analysis of selection in DNA repair genes was conducted upon consultation with Inigo Martincorena and Federico Abascal at the Wellcome Trust Sanger Institute and Santiago Gonzalez at IRB. All of the analyses described below were performed by me.

7.2 Widespread DNA repair defects in human cancer

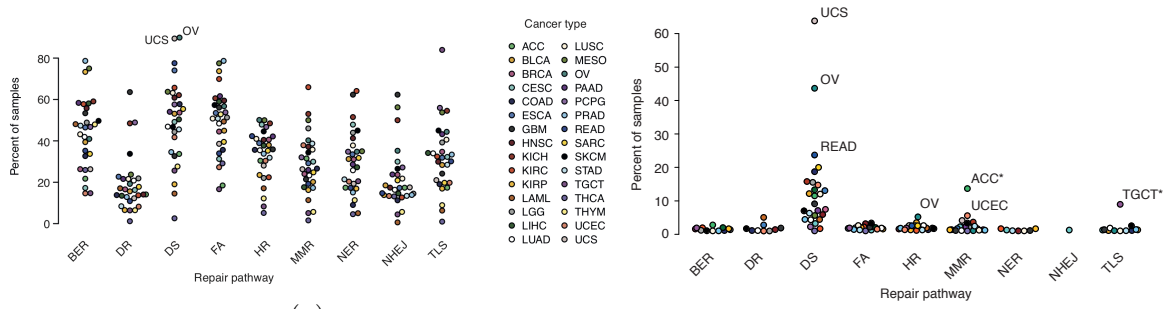
To leverage the concept of mutational spectra being shaped jointly by mutagen exposure and DNA repair status in human cancer genomes, we analysed the DNA repair

defects across 30 cancer types from TCGA and studied the appearance of various mutagenic processes in samples labelled as deficient or proficient in 9 DNA repair pathways: mismatch repair (MMR), base excision repair (BER), nucleotide excision repair (NER), homologous recombination repair (HRR), translesion synthesis (TLS), non-homologous end-joining (NHEJ), Fanconi anemia pathway (FA), direct repair (DR) and damage sensing (DS) pathways. In total, we analysed 81 core DNA repair genes across these pathways, and for a more in-depth analysis, we looked at additional 97 genes indirectly associated with DNA repair performed via these pathways (Knijnenburg et al. 2018, Pearl et al. 2015) (Appendix C). To assess the sample's DNA repair status, we annotated missense and loss-of-function mutations in these genes across 9,946 samples available from the TCGA collection.

If not stated otherwise, data was obtained from GDC (`https://cancergenome.nih.gov`) and filtered according to (Martincorena et al. 2017). Somatic and germline mutations were acquired using cgpCaVEMan (`http://cancerit.github.io/CaVEMan/`) and cgpPindel (`https://github.com/cancerit/cgpPindel`) variant callers. Copy number profiles, purities and ploidies of TCGA samples were estimated using ASCAT (Van Loo et al. 2010).

7.2.1 Monoallelic vs biallelic

To distinguish between the effects of complete or partial inactivation of DNA repair components, we classified samples with DNA repair defects into those with homozygous mutations (meaning biallelic inactivation) or heterozygous (monoallelic inactivation). Complete inactivation was considered to be a consequence of either a deep deletion, or two independent mutations, or a combination of mutation and gene silencing (Lahtz and Pfeifer 2011).



(a) (b)

Figure 7.1: (a) Monoallelic and (b) bi-allelic DNA repair pathway defects across TCGA. Each dot represents the percentage of samples with a certain defect within the respective cancer type. Corresponds to Supplementary Figure 6A,B in Volkova et al. 2019.

In line with the previous reports, we found that monoallelic defects in DNA repair were quite common - on average, over 50% of samples would have at least one gene of a given repair pathway mutated (Knijnenburg et al. 2018) (Figure 7.1a). In contrast, our analysis of biallelic inactivation events showed that these are rare (Figure 7.1b).

A notable exception is the damage sensing pathway, which is biallelically inactivated in over 11% of all samples due to mutations in commonly mutated DNA damage response genes such as *TP53* tumour suppressor gene. Amongst other DNA repair genes, most common is the biallelic inactivation of MMR (13% of adrenocortical carcinomas and 7% of endometrial cancers), as well as complete inactivation of TLS polymerase REV3L in testicular cancers (TGCT). Genes involved in direct damage reversal are impaired in 6% of leukaemias, and genes involved in homologous recombination repair occur to be inactivated in 6% of ovarian cancers.

7.2.2 Effect on mutation burden and spectra

Given the abundance of mono- and bi-allelic defects in DNA repair genes across all cancer types, we aimed to investigate how much of effect do they have on the mutational burden and mutational signatures present in respective tumours. To do that, we looked at the change in mutation burden per year between wild-type samples, samples with monoallelic inactivation, and samples with biallelic inactivation (if available) of the core genes in a given pathway. Here we also included defects in the proofreading domain of replicative polymerase ε ($POLE^{exo}$) as a separate pathway, as it is known to lead to hypermutation (Roberts and Gordenin 2014). We observed that an elevation in the number of mutations was not common - only 22 out of 300 combinations of cancer types showed a significant change in the mutational burden as shown by Wilcoxon rank-sum test (FDR 5%).

The most of the significant interactions associated with an increase in mutation burden were produced by defects in the damage sensing pathway and mismatch repair, as well as $POLE^{exo}$ defects (Figure 7.2). However, we did see some associations with direct repair and homologous recombination repair pathways.

Association between mutations in DNA repair genes and mutation burden could be caused simply by the fact that a higher total number of mutations means a higher chance of hitting the respective genes with mutations. Hence, we performed simulations to establish whether it was likely. Based on the distribution of synonymous mutations across the exome of all samples within certain cancer types, we simulated the number of mutations corresponding across core components of 9 DNA repair pathways such that it would produce the same number of synonymous mutations as observed, and tested whether the

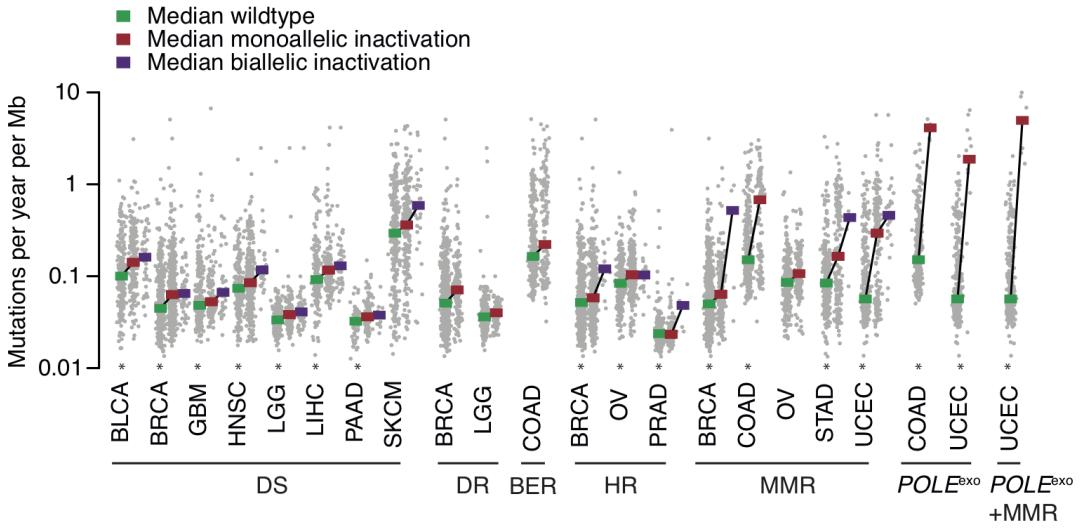


Figure 7.2: Changes in total number of mutations / year associated with DNA repair defects. Stars denote the cases where the presence of DNA repair defect is unlikely to be caused simply by mutation burden elevation. Corresponds to Supplementary Figure 6C in Volkova et al. 2019.

observed number of non-synonymous mutations in these genes across the set corresponds to expected. Using this approach, we identified that only about 70% of the associations we considered important could not be caused simply by the increase in mutation burden. Among these, damage sensing pathway, $POLE^{exo}$, HR and MMR deficiencies remained significant and showed a substantial increase in mutation burden in response to damaging mutations in respective pathways (Figure 7.2, stars).

Furthermore, we also assessed the mutational spectra of samples with and without DNA repair pathway defects. Upon subtraction of mutations caused by endogenous or non-DNA repair associated processes, such as signature SBS1 (associated with spontaneous deamination of 5meC) and APOBEC signatures (SBS2 and SBS13), we calculated cosine distances between the median mutational spectra of samples without mutations in a given pathway and those in the group with heterozygous mutations, and those in the group with homozygous mutations (if those groups of samples consisted of at least 4 samples with more than 100 mutations, to ensure the meaningfulness of the analysis of mutation spectra).

This comparison highlighted a striking difference in the profiles for proofreading domain mutations of POLE, for MMR mutations in stomach and uterine cancers, for DR mutations in glioblastomas and HR mutations in breast and ovarian cancers (Figure 7.3). These findings are in line with existence of strong mutational signatures associated with all these processes (Alexandrov et al. 2013b, Helleday, Eshtad, and Nik-Zainal 2014), apart from DR defects in GBM: a change in mutation spectrum of these samples is actu-

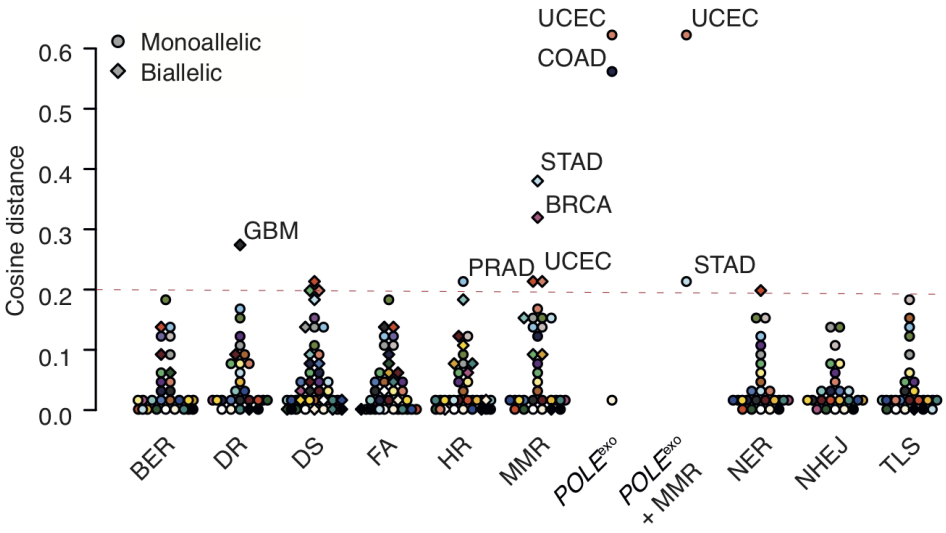


Figure 7.3: Changes in mutational spectra associated with DNA repair defects. Round points correspond to monoallelic defects, diamond-shaped points - to biallelic. Corresponds to Supplementary Figure 6D in Volkova et al. 2019.

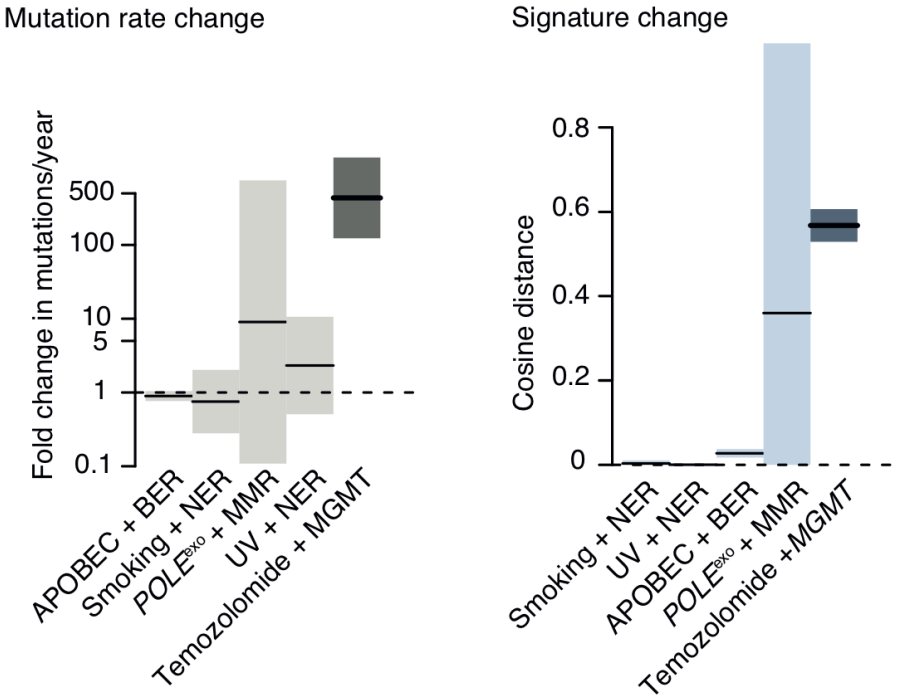
ally caused by the presence of strong treatment signature (temozolomide) in the samples defective in direct repair enzyme MGMT as compared to the rest of samples.

7.3 Damage-repair interactions

The observation that strong effects of DNA repair deficiency on the mutational burden and spectra are rare squares well with our experimental mutagenesis screen in which many DNA repair deficiencies displayed relatively mild mutational phenotypes under physiological conditions (Figure 3.2). However, several DNA repair-deficient genetic backgrounds yielded measurable genotoxin-repair interactions (Figure 6.6). Hence, to explore the scope and diversity of such effects in cancer, we adapted the negative-binomial model used to quantify signature changes in our mutagenesis screen to simultaneously extract signatures reflecting generally unknown genotoxic exposures and DNA repair effects thereon for a range of cancers with suspected genotoxic exposures (Section 2.4).

7.3.1 Confirmed interactions for temozolomide, $POLE^{exo}$ and APOBEC

Overall, the observed DNA repair and damage interaction effects were moderate, usually following the expected distributions (Figure 7.4). However, a number of noteworthy examples exist, shedding light on DNA damage-repair interactions also moulding cancer



(a) (b)

Figure 7.4: Interaction effects in human cancer. Fold-changes (a) in mutation rate and dissimilarity (b) between mutational signatures associated with repair-damage interactions. Corresponds to Figure 5A in Volkova et al. 2019.

genomes.

Temozolomide and MGMT

The strongest interaction, which is also therapeutically exploited, occurred between the human *agt-1* ortholog MGMT and alkylating chemotherapy drug temozolomide (TMZ) in temozolomide-treated glioblastomas (Figure 7.5) (Kim et al. 2015). To quantify it, we took the MGMT methylation data as well as the somatic mutations from exomes of glioblastoma multiformae (GBM) samples treated with TMZ (Kim et al. 2015). Using the model discussed in Section 2.4, we estimated the effect of MGMT methylation on one of the signatures present in glioblastoma exomes coming from 17 samples treated with TMZ: 11 with wild-type *MGMT* and 6 with *MGMT* hypermethylation. The model showed the best results for two signatures, one similar to COSMIC signature SBS1 and another one flat (upper panel of Figure 7.5), and suggested that the estimated change of the second signature was greater than 100-fold.

This is in good agreement with our experimental findings that the nature of mutation spectra detected upon EMS, DMS or MMS alkylation depends on the status of *agt-1*

(Figure 6.2b). We note that the signature associated with temozolomide exposure in MGMT deficient cancers leads to a more characteristic C>T spectrum in an NpCpY context (Y = C or T).

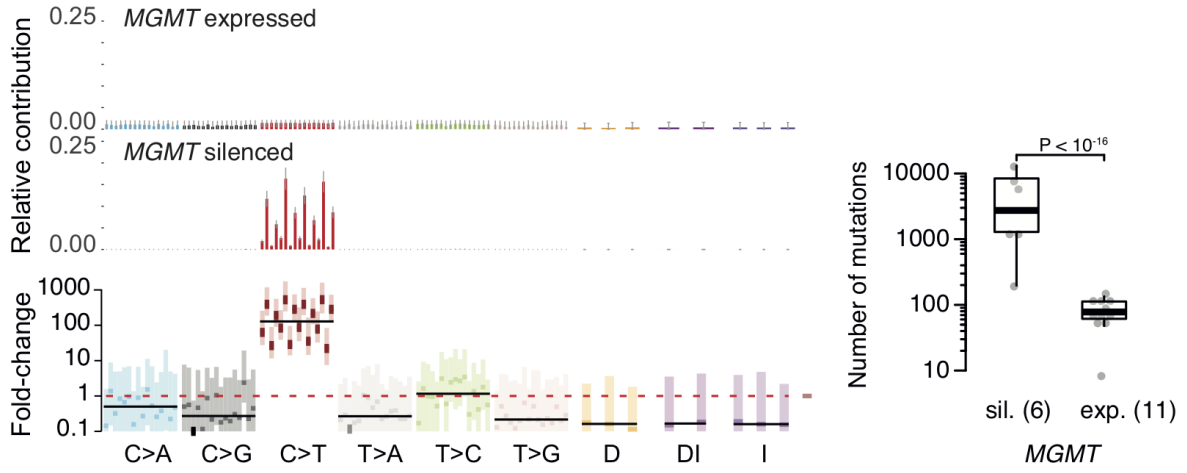


Figure 7.5: Interaction between temozolomide treatment and MGMT status. Top panel represents signature 2 in samples without *MGMT* silencing, middle panel - in samples with silenced *MGMT*. Lower panel represents the fold-change between these signatures. Boxplot on the right describes the number of mutations assigned to this signature in samples with and without *MGMT* silencing. Corresponds to Figure 5B in Volkova et al. 2019.

Concurrent mismatch repair deficiency and $POLE^{exo}$ defects

A further noteworthy interaction arose between MMR deficiency and a range of other mutational processes. These include concurrent POLE proofreading domain mutations manifesting in an increased C>A mutagenesis in a NpCpT context (Figure 7.6), as noted previously (Haradhvala et al. 2018, Shlien et al. 2015).

The effects of interaction between POLE and MMR defects were investigated using the uterine cancer cohort from TCGA (UCEC TCGA project) as indicated in (Haradhvala et al. 2018). Samples classified as MSI-H by Bethesda protocol (available in TCGA Clinical Explorer (Lee et al. 2015)) were considered to be MMR deficient, and samples with missense mutations in POLE proofreading domain (amino acids 267-472) were considered to have compromised proofreading activity.

Overall, we analysed 546 samples, 167 of which were labelled as MMR deficient, 55 - as having $POLE^{exo}$ defects, and 15 - as having both deficiencies. We identified five signatures, with a signature that demonstrated a 0.97 cosine similarity to COSMIC SBS10 signature (the one associated with POLE proofreading domain defects) being subject to interactions with different POLE mutations and POLE+MMR factor. In line with associ-

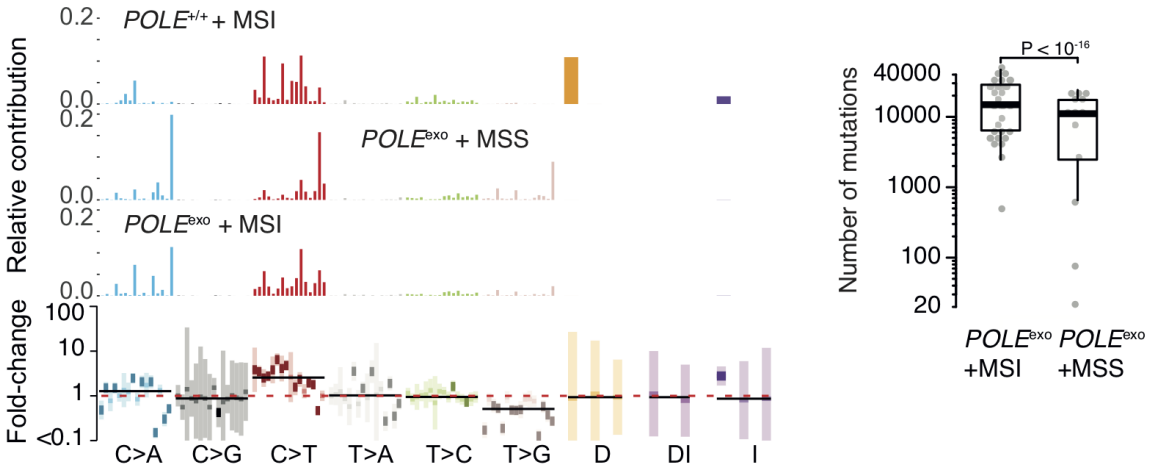


Figure 7.6: Signatures associated with MMR deficiency (first panel), and with $POLE^{exo}$ defects in MSI (second panel) and MSS (third panel) uterine cancers. Bottom panel represents the fold-change between the last two signatures, and the boxplot reflects the number of mutations assigned to $POLE^{exo}$ signature in MSI and MSS samples. Corresponds to Figure 5C in Volkova et al. 2019.

ations suggested previously, the transformation of POLE signature upon MMR deficiency closely resembled COSMIC signature SBS14, and signatures C1-2 from (Haradhvala et al. 2018) (Figure 7.6).

Concurrent mismatch repair deficiency across tissues

Moreover, MMR deficiency changes the rates of mutational indel processes, causing mononucleotide insertions and deletions in homopolymeric stretches (Figure 7.8), which arise at variable rate across tissues in MMR proficient samples (Alexandrov et al. 2018).

We screened all TCGA samples for defects in MMR genes *MLH1*, *PMS2*, *MSH2*, *MSH3* and *MSH6* (we excluded *MLH3* for the lack of samples with just the *MLH3* mutation and visible MMR deficiency phenotype). Samples were labelled as having MMR defects if they had a monoallelic/biallelic defect in MMR genes (assessed as described above) or were clinically determined to have MSI-H status as per TCGA Clinical Explorer (Lee et al. 2015). The samples were further filtered to exclude those with less than 100 mutations or more than 20000 mutations, and those whose mutational profile was dominated by a strong mutagenic process: APOBEC, POLE, UV or tobacco (i.e. we excluded those with cosine similarity higher than 0.7 to the respective signatures). From the remaining samples, we selected data from 8 tissue types: breast (BRCA), cervical (CESC), head and neck (HNSC), stomach (STAD), colorectal (COAD and READ), liver (LIHC), lung (LUAD and LUSC), prostate (PRAD), and uterine cancers (UCEC).

Using this dataset of 782 samples, we looked at the tissue effects (using colorectal as

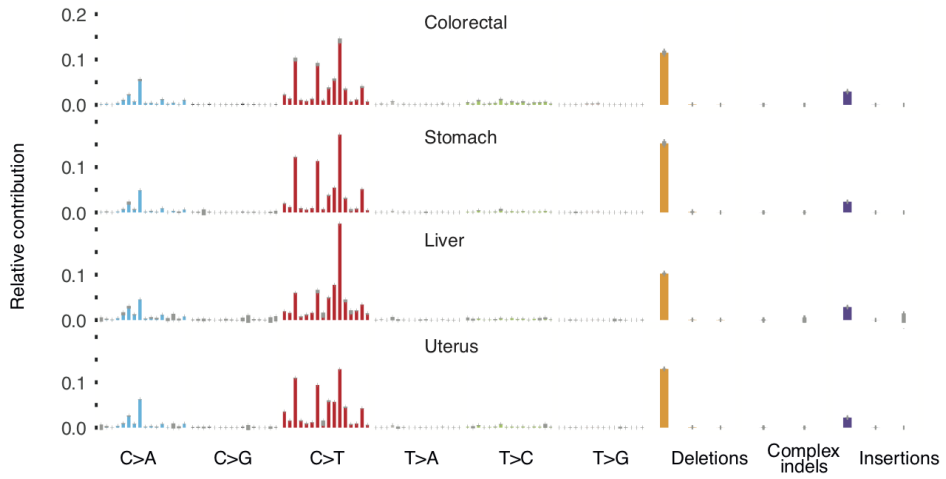


Figure 7.7: Signatures of MMR deficiency across different tissues. Corresponds to Supplementary Figure 7D in Volkova et al. 2019.

default and fitting 7 interaction effects for tissues - breast, cervix, head and neck, lung, stomach, uterus, liver), and noticed a change in the relative contribution of indels and CpG>TpG mutations in the MMR signature across some of the tissue types (Figure 7.7). When we compared age-adjusted rates of C>T mutations at CpG sites, and single-base deletions and insertions across samples from different cancer types with defected and wild-type MMR genes, we noticed a definite trend (Figure 7.8): the rates of these mutations were different across tissues, but the rate fold-changes in the respective mutation types between MMR deficient and proficient samples were similar across cancer types.

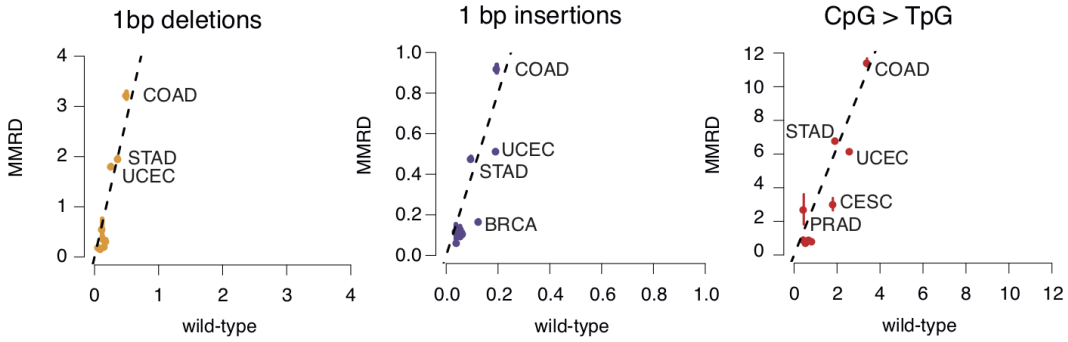


Figure 7.8: Indel and N[C>T]G rates in MMR-proficient and deficient cancer samples from different tissue types. Corresponds to Supplementary Figure 7D in Volkova et al. 2019.

Compared to MMR proficient samples, the rate of single-base deletions per year was on average 5-fold higher in MMR deficient samples following the same trend across different tissues; the rate of single-base insertions was on average 4-fold higher, and the rate of C>T at CpG sites was 3-fold higher in the samples with MMR defects compared to the

tissue baseline. These data also square well with the observation that MMR deficiency is associated with some additional mutational signatures of unknown aetiology (Meier et al. 2018), suggesting that these may reflect different mutagenic processes which are exacerbated by MMR deficiency.

APOBEC family of enzymes and translesion synthesis

APOBEC (apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like) family of enzymes represented another interesting example of a mutagenic process prone to alteration. APOBEC induces deamination of single-stranded cytosines, which is believed to be a key contributor of C>T and C>G mutations in a TpCpN context in a variety of cancer types. Uracil, the product of cytosine deamination, can pair with adenine, leading to C>T mutations upon replication; alternatively, uracil may be excised by uracil DNA glycosylase (UNG). It leaves an abasic site, which is thought to be replicated over by the error-prone TLS polymerase REV1, leading to C>G mutations (Morganella et al. 2016, Roberts and Gordenin 2014). Indeed, two mutational signatures involving either C>G and C>T mutations have been attributed to APOBEC and a lack of C>G mutations has been observed in a cancer cell line with UNG silencing (Kim et al. 2015, Petljak et al. 2019).

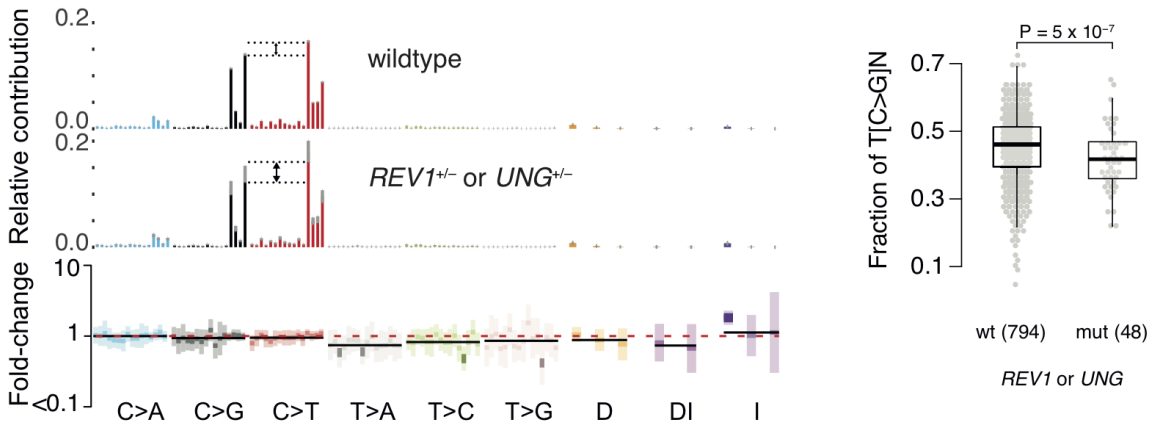
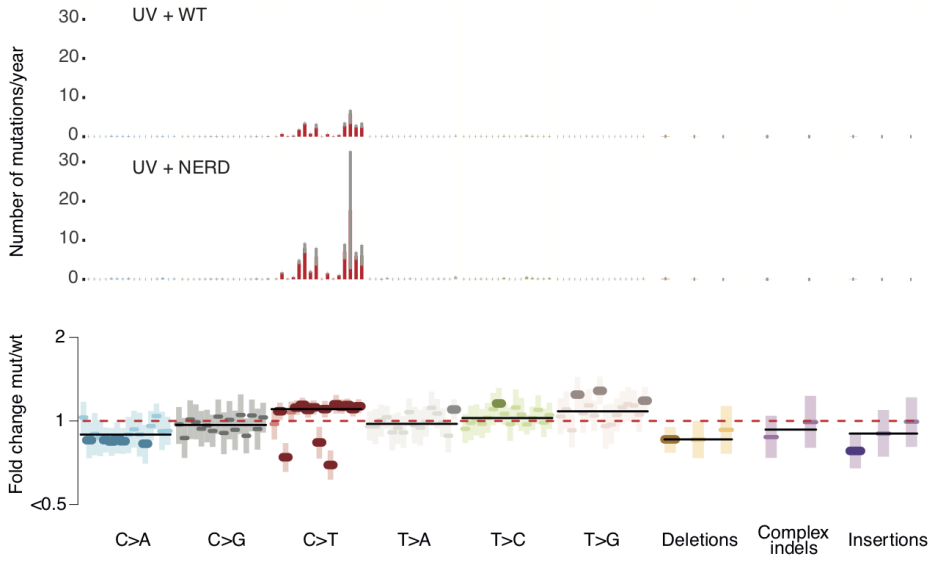


Figure 7.9: Change in APOBEC signature upon deficiency in REV1/UNG system along with the associated fold-change. The boxplot on the right reflects the fraction of T[C>G]N mutations compared to all mutations per sample in samples with wild-type and mutated *REV1* or *UNG*. Corresponds to Figure 5D in Volkova et al. 2019

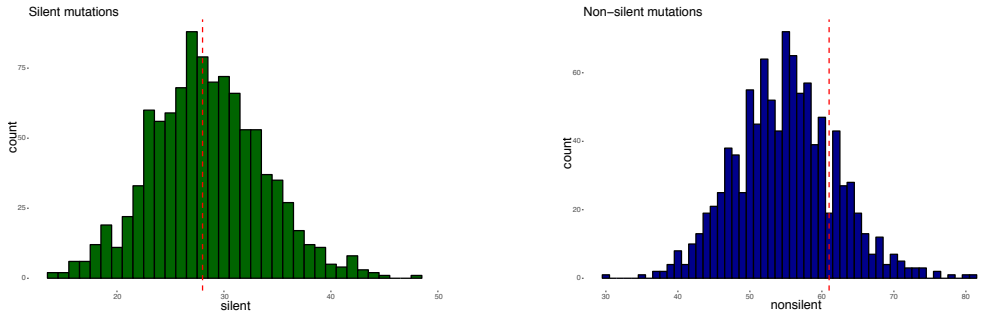
To assess the contributions of REV1 and UNG to APOBEC mutagenesis in primary cancers, we re-analysed 842 cancers with high APOBEC activity (the ones which showed cosine similarity of at least 0.8 to a combination of COSMIC signatures SBS2 and SBS13 and had between 50 and 15000 mutations per exome). We further stratified these samples into REV1/UNG wild-type (794 samples) and mutated group (48 samples with at least

monoallelic defects in REV1 or UNG). Analysis of APOBEC-specific T[C>A/C/T]N mutations showed an 8% decrease in the relative fraction of C>G mutations compared to C>T transversions in samples with defective REV1, indicating that translesion synthesis is indeed a critical contributor to APOBEC driven C>G mutagenesis (Figure 7.9).

7.3.2 No effects on mutagenesis for NER defects



(a)



(b) (c)

Figure 7.10: (a) Signature of UV-light in samples with wild-type NER genes (top) and those with defects in NER genes (middle panel), along with the fold-change (bottom panel). Corresponds to Supplementary Figure 7B in Volkova et al. 2019. (b-c) Simulations of silent (b) and non-silent (c) mutations within NER genes across SKCM dataset, showing that the observed values (dashed red line) fall into the distribution simulated based on the UV-associated spectrum of mutations and mutational burden per sample.

Notably, no significant effect was observed for NER variants in lung and skin cancers, although one might expect NER being involved in repairing bulky DNA adducts derived from tobacco smoke and UV light (Figure 7.4). Similarly, NER-deficient bladder cancers

with recurrent mutations in ERCC2/XPD did not show a strong effect of the mutation burden, despite reports of a mild increase of a signature similar to COSMIC signature 5 (Kim et al. 2016).

NER effects in UV-associated melanomas were estimated using samples from the TCGA cutaneous skin melanoma (SKCM) project (Cancer Genome Atlas Network 2015), where the similarity to a combination of COSMIC signatures SBS7a and SBS7b (previously associated with UV) was higher than 0.8. 397 patients were tested for having either a low expression of any of the NER genes in the tumour sample (below 20% of the median level in the dataset) or a somatic biallelic inactivation of a core NER gene in the tumour sample as described above. As the numbers were low, we also included samples with a high impact germline variant (homozygous or heterozygous) as predicted using Ensembl VEP (McLaren et al. 2016). The final set contained 9 samples labelled as NER defective.

We estimated the effects of having defects in NER pathway using a single signature. NER did not show a dramatic change in the profile of UV signature, only produced a 5% increase in C>T mutations, but led to an overall 2-fold increase in the mutation burden per year (Figure 7.10a). To check whether this is an effect of NER defects, or simply a consequence of increased mutational burden, we simulated UV-induced mutations in NER genes according to each sample's total burden of synonymous mutations and tested if the observed number of mutations in these genes across the dataset was in line with expected (Figure 7.10b). The observed number of nonsynonymous mutations across NER genes followed the distribution we generated via simulations (Figure 7.10c), which indicated that most of the signal was coming from the difference in the mutational burden. Together with the high variability of this fold-change effect, it led to the conclusion that there is no detectable effect of mutations in NER genes on the appearance of UV exposure signature.

Effects of DNA repair machinery defects on the signature of tobacco smoke were estimated using samples from LUAD and LUSC projects (Cancer Genome Atlas Research Network 2014b, Cancer Genome Atlas Research Network 2012). Out of 905 samples across the two datasets, we selected 219 samples with a high presence of COSMIC signature SBS4 (cosine similarity over 0.8), associated with tobacco smoking. Of these, 82 samples were labelled as being NER deficient (mono- or biallelically). Having extracted two signatures most similar to smoking and APOBEC associated signatures, we did not find a significant effect of NER defects on the exposure coefficient or mutational distribution of the signature associated with smoking (Figure 7.11).

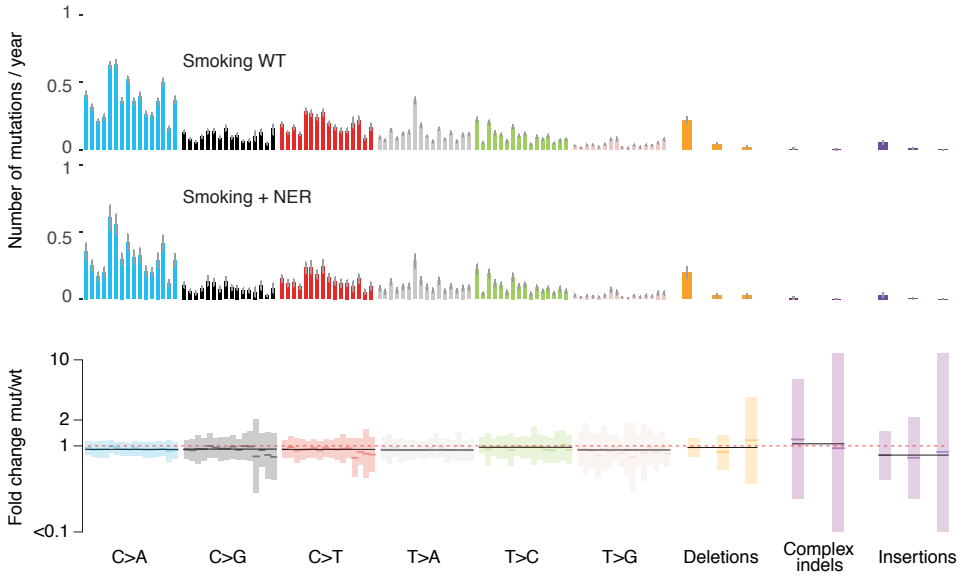


Figure 7.11: Change in tobacco signature upon NER defects. Corresponds to Supplementary Figure 7C in Volkova et al. 2019

7.4 DNA repair deficiency and somatic evolution of cancer

7.4.1 Selective pressure across DNA repair genes

A high level of recurrence of non-silent variants in a gene across cancer samples indicated that the given gene is positively selected and promotes cancer progression. To assess whether DNA repair mutations were positively or negatively selected in cancers, we calculated the dN/dS ratio of non-silent to silent mutations relative to its expected ratio based on silent variants and epigenetic covariates (Greenman et al. 2006, Martincorena et al. 2017). A value of $dN/dS = 1$ implies having the same ratio of non-silent to silent variants as expected, indicating no selection, while values greater than 1 signify positive selection, and negative values correspond to negative selection.

We analysed the dN/dS ratios of DNA repair genes of interest using the trinucleotide model according to (Martincorena et al. 2017). Background mutation rates were estimated using all genes except for consistently under-covered ones (Wang, Kim, and Chuang 2018). The significance of selection was assessed by comparing the relevant χ^2 statistic of the dN/dS ratio to χ^2-distribution with $df = 1$.

To study global selection across DNA repair pathways, we ran dNdSCV analysis separately within each cancer type using the samples with less than 1000 coding mutations. For gene-level analyses, we estimated the global and per-gene dN/dS values across all cancer types including all samples with less than 1000 coding mutations, using the extended

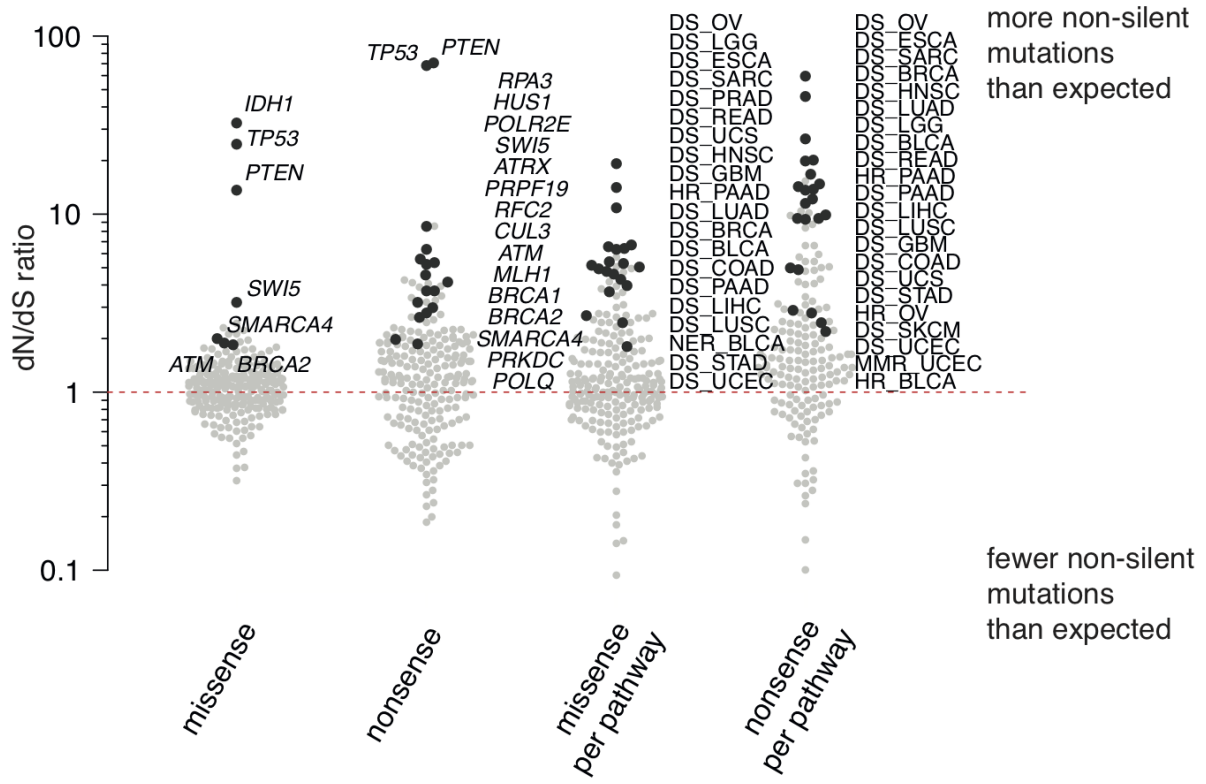


Figure 7.12: Selection across DNA repair pathways and individual genes. Corresponds to Figure 6A in Volkova et al. 2019.

set of 248 DNA repair associated genes (Appendix C). On the pathway level, we estimated the global dN/dS ratio over the list of core genes within each pathway in each cancer type. dN/dS ratios for missense and nonsense mutations were calculated separately. All of these ratios were then compared to 1 using χ^2-statistic of the squared z-score of $\log(dN/dS)$. Resulting p-values were adjusted for multiple testing within corresponding groups using the Benjamini-Hochberg procedure (Benjamini and Hochberg 1995).

Indeed, most of the DNA repair genes and pathways showed a high dN/dS ratio, indicating positive selection typical for driver gene mutations (Figure 7.12). The pathway-level analysis confirmed the pan-cancer excess of non-silent mutations in genes encoding for DNA damage sensing (DS) proteins including TP53, ATR and ATM (Table **??**). Equally, high impact mutations predominate in MMR genes in uterine and stomach cancers.

In single-gene analysis, 10/276 DNA repair genes displayed significant signs of positive selection in missense mutations, and 16/276 DNA repair genes were enriched in truncating mutations (FDR ¡ 10%; Figure 7.12, Table **??**). Among these the majority were known cancer genes such as *TP53*, *IDH1*, *PTEN*, *ATM*, *ATRX*, *SMARCA4*, *BRCA1/2*, *MLH1*, but we also detected new genes with mild signs of positive selection, namely the DNA damage sensing protein kinase *PRKDC* (dN/dS = 2 for nonsense variants; q-value =

0.01), the pre-mRNA processing factor *PRPF19*, involved in TC-NER, (dN/dS = 4.6 for nonsense variants; q-value = 0.005) and the homologous recombination repair gene *SWI5* (dN/dS = 3.2 for missense and 5.3 for nonsense variants; q-value = 0.05 and 0.07, respectively).

Negative selection for non-silent mutations is generally rare in cancer genomes and restricted to essential genes in haploid regions (Martincorena et al. 2017, Weghorn and Sunyaev 2017). However, missense mutations in *POLR2A* which encodes the DNA-directed RNA polymerase II subunit RPB1, showed a weak evidence of negative selection when analysed across samples with less than 500 mutations(dN/dS = 0.62, 95% CI $0.44 - 0.92$; q-value = 0.4), which was further exacerbated when analysed across the samples with a single copy of *POLR2A* (dN/dS = 0.26), suggesting that non-silent mutations in this gene are deleterious to cancer cell survival. Negative selection requires much greater statistical power to be detected compared to positive; hence, the q-value for this gene is high, and no other gene showed a clear indication of negative selection. However, negative selection in this gene is not surprising since *POLR2A* is an essential gene located on chromosome arm 17p, in cis with the tumour suppressor gene *TP53*. As 17p loss is common in cancers, this leaves affected cancer cells with only one copy of this essential gene in 20% of cancers rendering them sensitive to deleterious mutations. Thus, this genetic signal underscores the rationale that *POLR2A* might constitute a therapeutically exploitable vulnerability to α-amanitin in –17p/*TP53*-deficient cancers (Liu et al. 2015).

7.4.2 Relationship between mutation rate and cancer risk

The apparent discrepancy between the strong genetic evidence for DNA repair deficiency driving cancer both in inherited cancer syndromes, but also through somatically acquired mutations and the moderate phenotypes can, in fact, be explained from the perspective of somatic evolution. Given that mutations occur throughout a cancer patient's lifetime (Gerstung et al. 2017, Tomasetti, Vogelstein, and Parmigiani 2013), a DNA repair deficiency acquired at a late stage during tumour development may not have enough time to impact the mutational burden or signature substantially. In addition, it is important to consider the relationship between mutation rate and cancer risk, mindful that a number of driver mutations are needed for cancer transformation. Classic studies by (Armitage and Doll 1954) postulated that there exist about five rate-limiting steps needed for cancer formation, based on the observation that cancer risk increases approximately as the fifth power of age, a prediction in good agreement with recent estimates of 2-10 driver gene mutations in cancer genomes (Martincorena et al. 2017; Vogelstein and Kinzler 2015). As the chance to independently mutate multiple driver genes in the same cell becomes a

power of the mutation rate, a relatively small change in the mutation rate leads to a dramatic increase in the incidence rate of a cancer at any given age: Increasing the mutation rate by a factor of 2 leads to an approximately $2^5 = 32$ fold increased probability of 5 co-occurring mutations; clonal expansions reduce the exponent, but preserve the overall exponential scaling between mutation rate and cancer risk (Tomasetti et al. 2015).

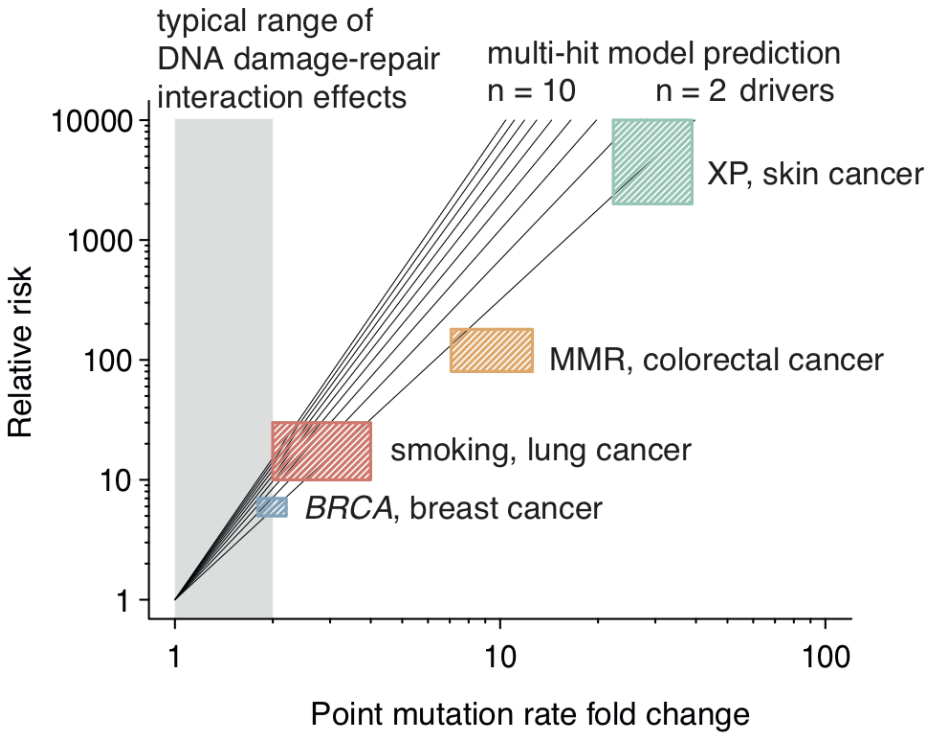


Figure 7.13: Relationship between fold-changes in cancer incidence and mutation rate for skin cancers and xeroderma pigmentosum (XP), colorectal cancers and MMR defects, lung cancer and smoking, breast cancer and smoking. The grey line denotes the region of expected mutation rate fold-change for NER gene defects considered in the analysis.

Indeed, the observed relation of relative risk and mutation rate increases in several cancer predisposition syndromes, but also smoking-related lung cancers follow this trend (Figure 7.13). As noted previously, MMR-deficient colorectal cancers have an 8-10 fold higher mutation burden than MMR-proficient carcinomas, but a ∼115 times higher burden in MMR-deficient patients. Similarly, HR-deficient breast cancers display a roughly 3-fold higher mutation burden, despite 20-40 fold increased risks for BRCA1/2 mutation carriers (Antoniou et al. 2003). Even more so, XP patients display a 100-fold increase of mutation rate (Zheng et al. 2014) but have an approximately 10,000 fold increased rate of skin cancers (Bradford et al. 2011). Under this model, high-impact SNPs in NER genes, which lead to a ∼ 2-fold increase in skin cancer risk (Wheless et al. 2012) would be expected to

cause a 20% increase in mutation rate, in agreement with our observations (Figure 7.13). While the exact number of driver gene mutations remains unknown in each cancer type, these data imply that small changes in mutation rates can have a large impact on cancer risk, and conversely noticeable risk factors may derive from rather moderate mutagenic effects.

Moderate effect sizes of genotoxin-repair interactions below a factor of 2 were found in our *C. elegans* screen in 25% of cases (Figure 6.6). Identifying small mutation rate changes in primary cancers proves difficult, especially as a DNA repair deficiency effect often only consists of an increase in mutation burden without any change to the mutational spectrum (Figure 7.2). Lastly, our analysis shows that the dN/dS ratio is around 2 for most DNA repair genes (Figure 7.12) implicating that only half of non-silent variants are truly cancer-causing, further diluting what may be a weak mutational signal.

7.5 Discussion

Our experimental data suggested that mutagenesis is fundamentally driven by the antagonism between DNA damage and repair; however, in the cancer genome, the manifestation of this principle may be masked by fluctuations in exposure doses, mutual timing of DNA repair deficiency and genotoxin exposure, as well as other confounding processes. Our cancer genome analysis showed that only the most extreme cases be reliably detected with the data currently available, and suggested that it is likely an issue of statistical power – the effects which would be expected based on evolutionary considerations have a magnitude similar to the current inter-personal variation, hence more genomes with longer exposures and tumour development times would be necessary to observe the actual diversity of damage-repair interactions. Nevertheless, there is a general selection process favouring mutations that inactivate DNA repair and DNA damage response genes during carcinogenesis but do not necessarily have a measurable effect on the mutational burden or mutational spectrum.

Indeed, while non-silent mutations in DNA repair genes are common, bi-allelic inactivation of both copies, which is generally required for loss of function, appears to be a relatively rare phenomenon. Frequent lack of strong mutational signature may appear confusing, however it becomes less surprising in light of timing and evolutionary considerations: to accumulate a substantial amount of mutations and manifest in a measurably increased mutation rate or altered spectrum, the DNA repair deficiency has to happen early enough to be in place when a mutagenic process starts to introduce damage to the DNA.

The notion of DNA damage and repair interactions shaping the mutational landscape

is critical from the perspective of DNA repair deficiency conditions because the same deficiency will yield a variety of spectra depending on the DNA damage (or replication errors) the genome experiences. Knowing the way how DNA repair deficiencies affect the mutational signatures of mutagens can also inform diagnostics and treatment choice, as it presents additional means of detecting clinically relevant DNA repair deficiencies, quantifying their onset timing and fraction of the tumour carrying this deficiency. Many DNA repair deficiencies indicate higher sensitivity to particular therapies, including PARP inhibition in case of HR deficiency, temozolomide treatment for DR-deficient glioblastomas, immunotherapy for MMR-deficient tumours, and platinum agent treatment for NER-deficient tumours (Hosoya and Miyagawa 2014).

Thus, to ascertain the status of a DNA repair pathway, a search for potential genetic and epigenetic defects in the DNA repair genes should be accompanied by a careful mutational signature analysis. Each of these parts alone can fail to indicate a DNA repair deficiency due to timing limitations or ambiguity in variant interpretation. Both of these limitations should be addressed by analysing larger datasets containing diverse variants across DNA repair-related genes and conducting experiments to validate the presence of functional consequences of DNA repair gene inactivation.

Chapter 8

Discussion

8.1 Summary of the main findings

In this thesis, I aimed to characterise a range of experimental mutational signatures induced by DNA repair deficiencies and genotoxins and to explore and quantify the factors affecting variation in mutational spectra. These goals were achieved by applying statistical modelling and investigating the genomic features of DNA damage and repair in nearly 3,000 samples from *C. elegans* and 10,000 cancer exomes from TCGA.

In brief, the results of the thesis are:

1. A catalogue of experimental mutational signatures of DNA repair deficiencies and genotoxin exposures. We described the genome-wide mutational spectra accumulated in 70 *C. elegans* lines with different DNA repair knockouts. Our analysis demonstrated the importance of TLS polymerases for the genome stability maintenance upon propagation over generations, suggested that defects in HRR components participating in different stages of HRR yield different mutations, characterised structural variation in the absence of crosslink repair, and quantified the protective capacity of apoptosis regulators. In addition, we showed that clustered mutations in the absence of specific exposures are almost exclusively arising due to mutagenic end-joining repair.

2. A high-resolution picture of the mutagenesis induced by MMR deficiency in *C. elegans* and gastrointestinal cancers. The comparison between the experimental and computational signatures helped to resolve the mixture between non-canonical MMR repair of mismatches produced by the deamination of 5-meC as well as indicates that other MMR signatures observed in cancers are likely to stem from the interactions of some mutational processes.

3. Characterisation of the experimental signatures of 12 genotoxins and comparison between genotoxin-induced mutagenesis in *C. elegans*, human iPS cell lines and cancer. While not being fully comparable, experimental and computational signatures demonstrated important similarities and discrepancies indicating the scopes of applicability of these systems.

4. A systematic and comprehensive analysis of how the interactions between mutagens and inactivations of DNA repair pathways shape the rate and spectra of mutations demonstrating that a third of all combinations of DNA repair deficiency and genotoxin exposure yield a change in the mutation rate or the spectrum.

5. Describing the range and mutational effects of DNA repair defects in cancer. Apart from confirming previously described interactions, we were able to quantify the reduction in C>G mutations induced by APOBEC in the absence of REV1 or UNG and demonstrate the tissue-specific differences in rates of base substitutions and indels caused by MMR deficiency. The distribution of damage-repair interaction effects observed in cancer was similar to the one observed in the experimental work: there are very few strong interactions, while the majority demonstrate weak effects. However, evolutionary considerations suggested that even small changes in mutation rates can lead to a steady increase in cancer incidence.

In general, this work demonstrated how a large-scale mutagenesis screen in a model system could inform and help to disentangle the components contributing to mutagenesis in cancer. More importantly, we also described the variability of mutational spectra of mutagenic processes across different genetic backgrounds, which demonstrated the prevalence of interaction effects between DNA repair components and damaging agents. Methodologically, this work presents two models for signature analysis in a controlled and uncontrolled environment incorporating the interactions of factors and allowing quantification of their contributions.

8.2 Conclusions

The analysis we conducted has several important implications. A comprehensive catalogue of high-resolution profiles for a wide range of genetic knockouts and mutagens will be a valuable resource for both *C. elegans* DNA repair biology and toxicology research and cancer genomics investigations as it presents the first set of experimentally derived interaction spectra.

First of all, our analysis demonstrated that *C. elegans* this a suitable model system to study genome-wide mutagenesis patterns and inform cancer research. The dramatic

difference in the magnitude of interaction effects observed in the screen and cancers also indicates the power of *C. elegans* model system over the aggregation of cancer data with unknown genotoxic exposure and timing of repair deficiencies. Additionally, the simplicity of the experimental setup allowed us to conduct a large-scale experiment covering many possible combinations of factors which would be tedious to perform in human cell lines, where only the analyses of either genotoxic or genetic factors were conducted up to date (Zou et al. 2018, Krüger and Piro 2019). Nevertheless, following on from our study, we expect that analysing DNA repair-defective model organisms and human cell lines, alone or in conjunction with defined genotoxic agents, will contribute to a more precise definition of mutational signatures occurring in cancer genomes and to establishing the aetiology of these signatures.

Demonstration of the complex relationship between mutagenic processes and DNA repair pathway status has the potential to change one of the main assumptions behind the mutational signature analysis stating the stability of signatures. The central insight of the interaction analysis is that a surprising 30% of experiments combining DNA damage and DNA repair deficiency demonstrated altered mutagenesis manifesting via a change in mutation rate or mutational spectrum. It is a correction to the widespread expectation in cancer genomics that there should be a one-to-one correspondence between the mutational signatures found in human cancers and a single genetic or genotoxic cause.

The high frequency yet the low impact of DNA repair defects observed in the TCGA collection of cancer exomes suggests that except for rare and extreme cases, the effect of these mutations appear small compared to the observed inter-sample heterogeneity. The presence of positive selection in several pathways across different cancer types, however, suggests that they can drive carcinogenesis without causing a substantial change in mutagenicity. This contradiction can be explained by the fact that the process of carcinogenesis usually requires a set of 2-10 driver gene mutations to occur in the right order (Tomasetti et al. 2015, Martincorena et al. 2018, Gerstung et al. 2018), with the chances of each next mutation rising steeply even for a small increase in mutation rates. Hence, small changes in mutation rate or pathogenicity of mutations, despite being undetectable due to technical or statistical limitations, can have a high impact on cancer development.

The analysis of mutational signatures has drawn much interest recently and dramatically extended our knowledge of the range and types of mutation patterns observed in human cancer and healthy tissues. However, one needs to recognise the variable nature of mutational spectra caused by the underlying dichotomy of damage and repair. This variability stems from the large difference between the amount of damage and amount of observed mutations which is defined by the efficiency and redundancy of DNA repair processes.

From the clinical point of view, our results may seem to diminish the value of mutational signature analysis. However, it instead suggests that simply associating signatures with genetic and clinical variables post-hoc is not enough, and some additional information about samples should be included to inform the signature analysis. A first attempt to employ clinical information to improve signature fitting was made by Robinson, Sharan, and Leiserson 2019 via tying the signature exposures to sample properties. The model suggested in this thesis aimed to accommodate for a spectrum change in a mutational signature attributable to these additional factors.

Overall, the insights described in this thesis bind together to deliver one important idea: a spectrum of mutations which can be observed upon sequencing is only a final sum of several damaging and repair processes acting with different signs. Hence, both these sides have to be taken into account when studying mutations in cancer or any other system. The actual range of potential interactions affecting the mutational spectra in cancer is still to be explored. Thus, one has to acknowledge this additional variation and integrate it into mutational signature analysis to ensure reliable interpretation and applicability in clinical oncology. Otherwise, unsupervised extraction of mutational signatures is likely only ever to represent the most striking exemplars of specific mutagenic constellations.

Lastly, the success of cross-species comparison shows that the fundamental laws of mutagenesis are acting in the same way across eukaryotic organisms from nematodes to humans. It reminds us that many mutational processes considered responsible for tumorigenesis are, in fact, not exclusive to cancer: these are the same forces as the ones driving variation and evolution of species.

8.3 Limitations of the analysis and potential improvements

The mutagenesis screen in *C. elegans* revealed a vast amount of details about the underlying mechanisms and provided the opportunity to quantify the mutagenic contributions and interactions of DNA damage and repair. However, like any other model system, *C. elegans* has some disadvantages which affect the scale of the alterations which can be observed.

The self-fertilising mode and high coding fraction of *C. elegans* genome do not allow it to accumulate as many mutations as a typical tumour. An average sample in the screen only carried 0.5 mutations per Mbps, whereas a typical tumour would have about 3-5 mutations per Mbps (as per ICGC March 2019 data release, `https://dcc.icgc.org`). Samples with the largest observed mutation rate – *pole-4; pms-2* double mutants –

could not be propagated beyond the 10-th generation, indicating that having more than 13-15 mutations/Mbps is lethal for *C. elegans* (although still normal for cancer). The experimental setup was adjusted to avoid as much negative selection as possible, but it still did not allow observing any genomic alterations which were too damaging to obtain a viable progeny. Moreover, aggregation of large-scale rearrangements was also limited as they tend to have a tremendous impact on the embryonic development of *C. elegans*. Therefore, it was not possible to observe as many rearrangements as typically observed in cancers.

Analysis of human data also posed its challenges. Interpersonal variability creates much noise in the data. The information which could account for some of it – data on the relevant exposures and age – was missing for many samples. Estimation of the interaction effects was further complicated by the lack of an accurate estimate of the mutual timing of the exposures and DNA deficiency onset as well as the clonal heterogeneity within the tumour, as both these factors define the amount of interaction-associated mutations which can be confidently detected.

Hence, there are some suggestions which could improve the analysis but did not fit into the scope of the projects I worked on:

- Generating and sequencing more replicates. The majority of the experiments had three biological replicates to estimate the variability of mutational spectra. However, even after pulling the data from different experiments together and estimating the variance of the signatures, some of the most variable interactions did not produce a feasible estimate due to insufficient data.
- Controlling for batch effects. Genotoxin exposure experiments were performed in batches, and the genotoxin solution or irradiation exposure were set up separately for each batch. Some of the mutagens were problematic in exploitation and caused some discrepancies in dose response between the sets of experiments performed at different times. Introducing better batch design and a high number of wild-type controls would help to avoid this issue.
- Adjusting the set of genotoxins and the list of tested combinations in light of the cases with observed interactions and the frequency of DNA repair gene mutations in cancer. It may be beneficial to produce more samples with a focus on the interactions between TLS polymerases and various carcinogens or chemotherapy agents as well as study more double mutants defective in more than one pathway.
- Applying a more sophisticated algorithm for the identification of DNA damage-repair interaction effects in cancer. A model which could accommodate more noise as well as handle the correlation between the factors could potentially capture more

signal.

- Finally, incorporating additional considerations in the analysis of interactions in cancer. Using more clinical variables such as nationality and gender and taking into account the timing of mutations could yield better patient stratification, but would also require even more data than is available now.

8.4 Outlook and future research

The work described in this thesis was performed in the scope of a project conceived over four years ago. With the scientific progress accelerating dramatically over the last decade (introduction of CRISPR-Cas9 system for targeted genome editing, the development of numerous single-cell techniques, advances in accuracy, speed and quality of sequencing, just to list a few), there is no doubt the same project, if it started today, would have even a larger scale and could potentially employ different experimental and computational methods.

The introduction of human organoids as model systems (Blokzijl et al. 2016, Drost et al. 2017) and the establishment of genome-wide mutagenesis experiments in CRISPR-modified human iPSCs and cancer cell lines (Zou et al. 2018, Petljak et al. 2019) allows exploring the damage-repair interactions in a human-based system. Although these experiments seem to be the closest reflection of the mutagenesis in human tissues, they are highly labour- and resource-demanding and prone to limitations. Organoids still represent a simplified model of tissue as they lack the proper structure and full set of cell types; it is also difficult to make organoids reach full maturation, and they often can not be expanded for as long as need (Xu et al. 2018). iPS cells have a similar issue with maturity and are also not fully representative of different human cell types (Hockemeyer and Jaenisch 2016). Hence, the simplicity and efficiency of other model systems, *C. elegans* in particular, will remain advantageous for creating large-scale screens.

Taking into account these aspects, one can suggest several avenues of improvement and enrichment of the current analysis, which could be undertaken by the successive researchers working on the quantification of DNA repair and damage signatures.

As mentioned in the previous section, generating more samples using the potentially impactful interactions could enhance the dataset. In addition, conducting confirmatory experiments for the strongest interactions using human-based model systems could increase the potential of this dataset for cancer research. Combining DNA repair deficiency and genotoxin exposure in human cells would highlight the differences and similarities between the model systems which should be taken into account when translating the findings from one to another.

Despite the availability of a massive catalogue of cancer data, many combinations of exposures and DNA repair defects could not be found in a sufficient number of samples. Hence, analysis of the datasets such as the data on metastatic tumours by the Hartwig medical foundation (Priestley et al. 2018), or the future datasets containing the mutational spectra of carcinogen-induced cancers, or cancers and cancer cell lines after exposure to genotoxic agents could expand the range of detectable interactions and reveal some new mutational signatures stemming from these interactions. Especially having a snapshot of the mutations present in a tumour before and after treatment could substantially enrich the analysis of interactions in cancer.

To produce a better quantitative model of DNA damage induction, removal of damage via DNA repair, and introduction of mutations, a deeper understanding of damage and repair specificities is required. Strong interactions, such as the alteration of the MMS signature under *polk-1* deficiency, pose a question: where do both the original and altered spectrum stem from? Is any of them reflecting the real distribution of the damage along the genome or only the context preferences of the TLS polymerases and repair enzymes? Hence, one of the directions of exploration should be experimental work on directly measuring the distribution and rate of damage induced by genotoxins (Sykora et al. 2018) as well as then the activity and success rate of DNA repair (Gassman and Holton 2019, Azqueta et al. 2019). Integrating these data with the observed spectra of mutations will provide a full quantitative description of the mutational input of the DNA damaging agents and repair systems.

Our screen, as well as other studies (Lange, Takata, and Wood 2011), suggested that DNA polymerases often serve as the primary agents introducing mutations. From that perspective, another prospective avenue of research would be to characterise the high-resolution genome-wide mutational spectra generated by all human polymerases on different substrates, including both normal and damaged substrates. Such a study could help to refine the role of polymerases and their defects in cancer and disease as well as give a better picture of the background mutagenesis in healthy tissues. Additionally, a better resolution of the genome-wide patterns of background mutagenesis could be achieved by analysing an extensive collection of normal tissues using high-sensitivity sequencing techniques such as BotSeq (Hoang et al. 2016, Dou et al. 2018). Confidently quantifying the distribution of rare mutations creating the somatic mosaicism in healthy individuals will improve the current understanding of carcinogenesis and ageing.

The development of these research directions could provide the building blocks for forming an exhaustive and accurate model of mutation acquisition upon different conditions. The current trends in the biomedical field suggest that whole-genome sequencing will be more and more common for all kinds of diseases which involve mutagenic pro-

cesses, and the fundamental knowledge of mutagenesis enhanced by the understanding of damage-repair interactions will provide the means to interpret and utilise these mutational spectra for better treatment and prevention of human diseases.

Bibliography

Abou-Zied, Osama K, Ralph Jimenez, and Floyd E Romesberg (2001). "Tautomerization dynamics of a model base pair in DNA". In: *Journal of the American Chemical Society* 123.19, pp. 4613–4614.

Abyzov, Alexej et al. (2012). "Somatic copy number mosaicism in human skin revealed by induced pluripotent stem cells". In: *Nature* 492.7429, p. 438.

Akaike, Hirotogu (1992). "Information Theory and an Extension of the Maximum Likelihood Principle". In: *Springer Series in Statistics*, pp. 610–624.

Aksenova, Anna et al. (2010). "Mismatch repair-independent increase in spontaneous mutagenesis in yeast lacking non-essential subunits of DNA polymerase ϵ". en. In: *PLoS Genet.* 6.11, e1001209.

Alberts, Bruce et al. (2007). *Molecular Biology of the Cell*.

Albertson, Tina M et al. (2009). "DNA polymerase epsilon and delta proofreading suppress discrete mutator and cancer phenotypes in mice". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 106.40, pp. 17101–17104.

Alexandrov, Ludmil et al. (2018). "The Repertoire of Mutational Signatures in Human Cancer". In: *bioRxiv*.

Alexandrov, Ludmil B et al. (2013a). "Deciphering signatures of mutational processes operative in human cancer". en. In: *Cell Rep.* 3.1, pp. 246–259.

Alexandrov, Ludmil B et al. (2013b). "Signatures of mutational processes in human cancer". en. In: *Nature* 500.7463, pp. 415–421.

Alexandrov, Ludmil B et al. (2015). "Clock-like mutational processes in human somatic cells". en. In: *Nat. Genet.* 47.12, pp. 1402–1407.

Alexandrov, Ludmil B et al. (2016). "Mutational signatures associated with tobacco smoking in human cancer". en. In: *Science* 354.6312, pp. 618–622.

Alkodsi, Amjad et al. (2019). "Distinct subtypes of diffuse large B-cell lymphoma defined by hypermutated genes". In: *Leukemia*, p. 1.

Antoniou, A et al. (2003). "Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case Series unselected for family history: a combined analysis of 22 studies". en. In: *Am. J. Hum. Genet.* 72.5, pp. 1117–1130.

Antoshechkin, Igor and Paul W Sternberg (2007). "The versatile worm: genetic and genomic resources for Caenorhabditis elegans research". In: *Nature Reviews Genetics* 8.7, p. 518.

Arlt, Volker M, Marie Stiborova, and Heinz H Schmeiser (2002). "Aristolochic acid as a probable human cancer hazard in herbal remedies: a review". In: *Mutagenesis* 17.4, pp. 265–277.

Armitage, P and R Doll (1954). "The Age Distribution of Cancer and a Multi-stage Theory of Carcinogenesis". In: *British Journal of Cancer* 8.1, pp. 1–12.

Armstrong, Bruce K, Anne Kricker, and Dallas R English (1997). "Sun exposure and skin cancer". In: *Australasian journal of dermatology* 38.S1, S1–S6.

Attaluri, Sivaprasad et al. (2009). "DNA adducts of aristolochic acid II: total synthesis and site-specific mutagenesis studies in mammalian cells". In: *Nucleic acids research* 38.1, pp. 339–352.

Azqueta, Amaya et al. (2019). "DNA repair as a human biomonitoring tool; comet assay approaches". In: *Mutation Research/Reviews in Mutation Research*.

Baez-Ortega, Adrian and Kevin Gori (2017). "Computational approaches for discovery of mutational signatures in cancer". In: *Briefings in bioinformatics* 20.1, pp. 77–88.

Baross-Francis, Agnes et al. (2001). "Elevated mutant frequencies and increased C : G→T : A transitions in Mlh1/ versus Pms2/ murine small intestinal epithelial cells". In: *Oncogene* 20.5, pp. 619–625.

Baskar, Rajamanickam et al. (2012). "Cancer and radiation therapy: current advances and future directions". In: *International journal of medical sciences* 9.3, p. 193.

Behjati, Sam et al. (2014). "Genome sequencing of normal cells reveals developmental lineages and mutational processes". In: *Nature* 513.7518, p. 422.

Behjati, Sam et al. (2016). "Mutational signatures of ionizing radiation in second malignancies". en. In: *Nat. Commun.* 7, p. 12605.

Bellacosa, A (2001). "Functional interactions and signaling properties of mammalian DNA mismatch repair proteins". en. In: *Cell Death Differ.* 8.11, pp. 1076–1092.

Benjamini, Yoav and Yosef Hochberg (1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 57.1, pp. 289–300.

Bernstein, Carol et al. (2013). "DNA damage, DNA repair and cancer". In: *New Research Directions in DNA Repair*. IntechOpen.

Betancourt, Michael and Mark Girolami (2015). "Hamiltonian Monte Carlo for hierarchical models". In: *Current trends in Bayesian methodology with applications* 79, p. 30.

Bhargava, Ragini, David O Onyango, and Jeremy M Stark (2016). "Regulation of single-strand annealing and its role in genome maintenance". In: *Trends in Genetics* 32.9, pp. 566–575.

Blei, David M, John D Lafferty, et al. (2007). "A correlated topic model of science". In: *The Annals of Applied Statistics* 1.1, pp. 17–35.

Blei, David M, Andrew Y Ng, and Michael I Jordan (2003). "Latent dirichlet allocation". In: *Journal of machine Learning research* 3.Jan, pp. 993–1022.

Blokzijl, Francis et al. (2016). "Tissue-specific mutation accumulation in human adult stem cells during life". In: *Nature* 538.7624, p. 260.

Blokzijl, Francis et al. (2018). "MutationalPatterns: comprehensive genome-wide analysis of mutational processes". In: *Genome medicine* 10.1, p. 33.

Boer, Jan de and Jan HJ Hoeijmakers (2000). "Nucleotide excision repair and human syndromes". In: *Carcinogenesis* 21.3, pp. 453–460.

Boffetta, Paolo, Nadia Jourenkova, and Per Gustavsson (1997). "Cancer risk from occupational and environmental exposure to polycyclic aromatic hydrocarbons". In: *Cancer Causes & Control* 8.3, pp. 444–472.

Boissière-Michot, Florence et al. (2016). "Immunohistochemical staining for p16 and BRAFV600E is useful to distinguish between sporadic and hereditary (Lynch syndrome-related) microsatellite instable colorectal carcinomas". en. In: *Virchows Arch.* 469.2, pp. 135–144.

Boland, C R et al. (1998). "A National Cancer Institute Workshop on Microsatellite Instability for cancer detection and familial predisposition: development of international criteria for the determination of microsatellite instability in colorectal cancer". en. In: *Cancer Res.* 58.22, pp. 5248–5257.

Bonneville, Russell et al. (2017). "Landscape of Microsatellite Instability Across 39 Cancer Types". en. In: *JCO Precis Oncol* 2017.

Boot, Arnoud et al. (2018). "In-depth characterization of the cisplatin mutational signature in human cell lines and in esophageal and liver tumors". en. In: *Genome Res.* 28.5, pp. 654–665.

Boulton, Simon J (2010). "DNA repair: Decision at the break point". In: *Nature* 465.7296, p. 301.

Boulton, Simon J et al. (2004). "BRCA1/BARD1 orthologs required for DNA repair in Caenorhabditis elegans". In: *Current Biology* 14.1, pp. 33–39.

Boysen, Gunnar et al. (2009). "The formation and biological significance of N7-guanine adducts". In: *Mutation Research/Genetic Toxicology and Environmental Mutagenesis* 678.2, pp. 76–94.

Bradford, Porcia T et al. (2011). "Cancer and neurologic degeneration in xeroderma pigmentosum: long term follow-up characterises the role of DNA repair". en. In: *J. Med. Genet.* 48.3, pp. 168–176.

Brenner, Michaela and Vincent J Hearing (2008). "The protective role of melanin against UV damage in human skin". In: *Photochemistry and photobiology* 84.3, pp. 539–549.

Bronner, C E et al. (1994). "Mutation in the DNA mismatch repair gene homologue hMLH1 is associated with hereditary non-polyposis colon cancer". en. In: *Nature* 368.6468, pp. 258–261.

Brookes, Peter and Philip D Lawley (1961). "The reaction of mono-and di-functional alkylating agents with nucleic acids". In: *Biochemical Journal* 80.3, p. 496.

Brouwer, Judith R, Rob Willemsen, and Ben A Oostra (2009). "Microsatellite repeat instability and neurological disease". In: *Bioessays* 31.1, pp. 71–83.

Brunet, Jean-Philippe et al. (2004). "Metagenes and molecular pattern discovery using matrix factorization". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 101.12, pp. 4164–4169.

Bryan, D Suzi et al. (2014). "High resolution mapping of modified DNA nucleobases using excision repair enzymes". In: *Genome research* 24.9, pp. 1534–1542.

Buhard, Olivier et al. (2006). "Multipopulation analysis of polymorphisms in five mononucleotide repeats used to determine the microsatellite instability status of human tumors". en. In: *J. Clin. Oncol.* 24.2, pp. 241–251.

Cadet, Jean and Thierry Douki (2018). "Formation of UV-induced DNA damage contributing to skin cancer development". In: *Photochemical & Photobiological Sciences* 17.12, pp. 1816–1841.

Cadet, Jean et al. (2014). "One-electron oxidation reactions of purine and pyrimidine bases in cellular DNA". In: *International journal of radiation biology* 90.6, pp. 423–432.

Caldecott, Keith W (2008). "Single-strand break repair and genetic disease". In: *Nature Reviews Genetics* 9.8, p. 619.

Campbell, Peter J et al. (2017). "Pan-cancer analysis of whole genomes". In: *BioRxiv*, p. 162784.

Cancer Genome Atlas Network (2012). "Comprehensive molecular characterization of human colon and rectal cancer". en. In: *Nature* 487.7407, pp. 330–337.

— (2015). "Genomic Classification of Cutaneous Melanoma". en. In: *Cell* 161.7, pp. 1681–1696.

Cancer Genome Atlas Research Network (2012). "Comprehensive genomic characterization of squamous cell lung cancers". en. In: *Nature* 489.7417, pp. 519–525.

— (2014a). "Comprehensive molecular characterization of gastric adenocarcinoma". en. In: *Nature* 513.7517, pp. 202–209.

— (2014b). “Comprehensive molecular profiling of lung adenocarcinoma”. en. In: *Nature* 511.7511, pp. 543–550.

Cancer Genome Atlas Research Network et al. (2013). “Integrated genomic characterization of endometrial carcinoma”. en. In: *Nature* 497.7447, pp. 67–73.

Cemgil, Ali Taylan (2009). “Bayesian inference for nonnegative matrix factorisation models”. en. In: *Comput. Intell. Neurosci.* P. 785152.

Chadt, Jiri et al. (2008). “Monitoring of dimethyl sulphate-induced N3-methyladenine, N7-methylguanine and O6-methylguanine DNA adducts using reversed-phase high performance liquid chromatography and mass spectrometry”. In: *Journal of Chromatography B* 867.1, pp. 43–48.

Chahwan, Richard et al. (2012). “AIDing antibody diversity by error-prone mismatch repair”. In: *Seminars in immunology*. Vol. 24. 4. Elsevier, pp. 293–300.

Chakraborty, Anirban et al. (2016). “Classical non-homologous end-joining pathway utilizes nascent RNA for error-free double-strand break repair of transcribed genes”. In: *Nature communications* 7, p. 13049.

Chan, Kin and Dmitry A Gordenin (2015). “Clusters of multiple mutations: incidence and molecular mechanisms”. In: *Annual review of genetics* 49, pp. 243–267.

Chang, Howard HY et al. (2017). “Non-homologous DNA end joining and alternative pathways to double-strand break repair”. In: *Nature reviews Molecular cell biology* 18.8, p. 495.

Chen, Jian-Min, Claude Férec, and David N Cooper (2013). “Patterns and mutational signatures of tandem base substitutions causing human inherited disease”. In: *Human mutation* 34.8, pp. 1119–1130.

Christensen, Sharon et al. (2019). “5-Fluorouracil treatment induces characteristic T¿ G mutations in human cancer”. In: *bioRxiv*, p. 681262.

Ciccia, Alberto and Stephen J Elledge (2010). “The DNA damage response: making it safe to play with knives”. In: *Molecular cell* 40.2, pp. 179–204.

Cohen, Seth M and Stephen J Lippard (2001). “Cisplatin: from DNA damage to cancer chemotherapy”. In:

Comon, Pierre and Christian Jutten (2010). *Handbook of Blind Source Separation: Independent component analysis and applications*. Academic press.

Connor, Frances et al. (2018). “Mutational landscape of a chemically-induced mouse model of liver cancer”. In: *Journal of hepatology* 69.4, pp. 840–850.

Conrad, Donald F et al. (2011). “Variation in genome-wide mutation rates within and between human families”. In: *Nature genetics* 43.7, p. 712.

Cortes-Ciriano, Isidro et al. (2017). “A molecular portrait of microsatellite instability across multiple cancers”. en. In: *Nat. Commun.* 8, p. 15180.

Davidson, Philip R et al. (2017). "A pooled mutational analysis identifies ionizing radiation-associated mutational signatures conserved between mouse and human malignancies". In: *Scientific reports* 7.1, p. 7645.

Davies, Helen et al. (2017). "HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures". In: *Nature medicine* 23.4, p. 517.

Davis, Anthony J and David J Chen (2013). "DNA double strand break repair via non-homologous end-joining". In: *Translational cancer research* 2.3, p. 130.

De Bont, Rinne and Nik Van Larebeke (2004). "Endogenous DNA damage in humans: a review of quantitative data". In: *Mutagenesis* 19.3, pp. 169–185.

Deem, Angela et al. (2011). "Break-induced replication is highly inaccurate". In: *PLoS biology* 9.2, e1000594.

Degtyareva, Natalya P et al. (2019). "Mutational signatures of redox stress in yeast single-strand DNA and of aging in human mitochondrial DNA share a common feature". In: *PLoS biology* 17.5, e3000263.

Degtyareva, Natasha P et al. (2002). "Caenorhabditis elegans DNA mismatch repair gene msh-2 is required for microsatellite stability and maintenance of genome integrity". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 99.4, pp. 2158–2163.

Delaney, James C and John M Essigmann (2004). "Mutagenesis, genotoxicity, and repair of 1-methyladenine, 3-alkylcytosines, 1-methylguanine, and 3-methylthymine in alkB Escherichia coli". In: *Proceedings of the National Academy of Sciences* 101.39, pp. 14051–14056.

Denissenko, Mikhail F et al. (1998). "Slow repair of bulky DNA adducts along the non-transcribed strand of the human p53 gene may explain the strand bias of transversion mutations in cancers". In: *Oncogene* 16.10, p. 1241.

Denver, Dee R et al. (2004). "Abundance, distribution, and mutation rates of homopolymeric nucleotide runs in the genome of Caenorhabditis elegans". en. In: *J. Mol. Evol.* 58.5, pp. 584–595.

Denver, Dee R et al. (2005). "Mutation rates, spectra and hotspots in mismatch repair-deficient Caenorhabditis elegans". en. In: *Genetics* 170.1, pp. 107–113.

Denver, Dee R et al. (2009). "A genome-wide view of Caenorhabditis elegans base-substitution mutation processes". In: *Proceedings of the National Academy of Sciences* 106.38, pp. 16310–16314.

Devarajan, Karthik (2008). "Nonnegative matrix factorization: an analytical and interpretive tool in computational biology". In: *PLoS computational biology* 4.7, e1000029.

Devasagayam, TPA et al. (2004). "Free radicals and antioxidants in human health: current status and future prospects". In: *Japi* 52.794804, p. 4.

Ding, Li et al. (2008). “Somatic mutations affect key pathways in lung adenocarcinoma”. In: *Nature* 455.7216, p. 1069.

Dong, Hongbin et al. (2015). “Update of the human and mouse Fanconi anemia genes”. In: *Human genomics* 9.1, p. 32.

Dou, Yanmei et al. (2018). “Detecting somatic mutations in normal cells”. In: *Trends in Genetics* 34.7, pp. 545–557.

Dow, Michelle et al. (2018). “Integrative genomic analysis of mouse and human hepatocellular carcinoma”. In: *Proceedings of the National Academy of Sciences* 115.42, E9879–E9888.

Drabløs, Finn et al. (2004). “Alkylation damage in DNA and RNA—repair mechanisms and medical significance”. In: *DNA repair* 3.11, pp. 1389–1407.

Drost, Jarno et al. (2017). “Use of CRISPR-modified human stem cell organoids to study the origin of mutational signatures in cancer”. In: *Science* 358.6360, pp. 234–238.

Drummond, J T et al. (1995). “Isolation of an hMSH2-p160 heterodimer that restores DNA mismatch repair to tumor cells”. en. In: *Science* 268.5219, pp. 1909–1912.

Dudley, J C et al. (2016). “Microsatellite Instability as a Biomarker for PD-1 Blockade”. In: *Clinical Cancer Research* 22.4, pp. 813–820.

Durno, Carol A et al. (2015). “Phenotypic and genotypic characterisation of biallelic mismatch repair deficiency (BMMR-D) syndrome”. en. In: *Eur. J. Cancer* 51.8, pp. 977–983.

Engert, Andreas, Jurgen Wolf, and Volker Diehl (1999). “Treatment of advanced Hodgkin’s lymphoma: standard and experimental approaches.” In: *Seminars in hematology*. Vol. 36. 3, pp. 282–289.

Faili, Ahmad et al. (2002). “Induction of somatic hypermutation in immunoglobulin genes is dependent on DNA polymerase iota”. In: *Nature* 419.6910, p. 944.

Figueroa-González, Gabriela and Carlos Pérez-Plasencia (2017). “Strategies for the evaluation of DNA damage and repair mechanisms in cancer”. In: *Oncology letters* 13.6, pp. 3982–3988.

Fischer, Andrej et al. (2013). “EMu: probabilistic inference of mutational processes and their localization in the cancer genome”. In: *Genome biology* 14.4, R39.

Fishel, R et al. (1994). “The human mutator gene homolog MSH2 and its association with hereditary nonpolyposis colon cancer”. en. In: *Cell* 77.1, 1 p following 166.

Flibotte, Stephane et al. (2010). “Whole-genome profiling of mutagenesis in Caenorhabditis elegans”. en. In: *Genetics* 185.2, pp. 431–441.

Forbes, Simon A et al. (2015). “COSMIC: exploring the world’s knowledge of somatic mutations in human cancer”. In: *Nucleic Acids Research* 43.D1, pp. D805–D811.

Forment, Josep V, Abderrahmane Kaidi, and Stephen P Jackson (2012). "Chromothripsis and cancer: causes and consequences of chromosome shattering". In: *Nature Reviews Cancer* 12.10, p. 663.

Fredriksson, Nils Johan et al. (2017). "Recurrent promoter mutations in melanoma are defined by an extended context-specific mutational signature". In: *PLoS genetics* 13.5, e1006773.

Fronza, Gilberto and Barry Gold (2004). "The biological effects of N3-methyladenine". In: *Journal of cellular biochemistry* 91.2, pp. 250–257.

Funkhouser Jr, William K et al. (2012). "Relevance, pathogenesis, and testing algorithm for mismatch repair-defective colorectal carcinomas: a report of the association for molecular pathology". en. In: *J. Mol. Diagn.* 14.2, pp. 91–103.

Funnell, Tyler et al. (2019). "Integrated structural variation and point mutation signatures in cancer genomes using correlated topic models". In: *PLoS computational biology* 15.2, e1006799.

Gan, Gregory N et al. (2008). "DNA polymerase zeta (pol ζ) in higher eukaryotes". In: *Cell research* 18.1, p. 174.

Gandini, Sara et al. (2008). "Tobacco smoking and cancer: A meta-analysis". In: *International journal of cancer* 122.1, pp. 155–164.

Garry, Vincent F et al. (1979). "Ethylene oxide: evidence of human chromosomal effects". In: *Environmental mutagenesis* 1.4, pp. 375–382.

Gassman, Natalie R and Nathaniel W Holton (2019). "Targets for repair: detecting and quantifying DNA damage with fluorescence-based methodologies". In: *Current opinion in biotechnology* 55, pp. 30–35.

Genschel, J et al. (1998). "Isolation of MutSbeta from human cells and comparison of the mismatch repair specificities of MutSbeta and MutSalpha". en. In: *J. Biol. Chem.* 273.31, pp. 19895–19901.

Georgieva, Daniela et al. (2019). "Detection of Base Analogs Incorporated During DNA Replication by Nanopore Sequencing". In: *bioRxiv*, p. 549220.

Gerber, Christoph and Heinz-Gerhard Toelle (2009). "What happened: The chemistry side of the incident with EMS contamination in Viracept tablets". In: *Toxicology letters* 190.3, pp. 248–253.

Gerstung, Moritz et al. (2017). "The evolutionary history of 2,658 cancers". In: *bioRxiv*.

Gerstung, Moritz et al. (2018). "The evolutionary history of 2,658 cancers". In: *BioRxiv*, p. 161562.

Giglia-Mari, Giuseppina, Angelika Zotter, and Wim Vermeulen (2011). "DNA damage response". In: *Cold Spring Harbor perspectives in biology* 3.1, a000745.

Goellner, Eva M, Christopher D Putnam, and Richard D Kolodner (2015). "Exonuclease 1-dependent and independent mismatch repair". en. In: *DNA Repair* 32, pp. 24–32.

Golding, Nick (2018). *greta: Simple and Scalable Statistical Modelling in R.*

Gori, Kevin and Adrian Baez-Ortega (2018). "sigfit: flexible Bayesian inference of mutational signatures". In: *bioRxiv*, p. 372896.

Gradia, S et al. (1999). "hMSH2-hMSH6 forms a hydrolysis-independent sliding clamp on mismatched DNA". en. In: *Mol. Cell* 3.2, pp. 255–261.

Grasso, Francesca and Teresa Frisan (2015). "Bacterial genotoxins: merging the DNA damage response into infection biology". In: *Biomolecules* 5.3, pp. 1762–1782.

Greaves, Mel and Carlo C Maley (2012). "Clonal evolution in cancer". In: *Nature* 481.7381, p. 306.

Green, Adèle C et al. (2011). "Reduced melanoma after regular sunscreen use: randomized trial follow-up". In: *Journal of clinical oncology* 29.3, pp. 257–263.

Greenblatt, Marc S et al. (2001). "TP53 mutations in breast cancer associated with BRCA1 or BRCA2 germ-line mutations: distinctive spectrum and structural distribution". In: *Cancer research* 61.10, pp. 4092–4097.

Greenman, Chris et al. (2006). "Statistical analysis of pathogenicity of somatic mutations in cancer". In: *Genetics* 173.4, pp. 2187–2198.

Greer, Eric Lieberman et al. (2015). "DNA Methylation on N6-Adenine in C. elegans". In: *Cell* 161.4, pp. 868–878.

Gregory, T Ryan (2005). "Synergy between sequence and size in large-scale genomics". In: *Nature Reviews Genetics* 6.9, p. 699.

Grin, Inga and Alexander A Ishchenko (2016). "An interplay of the base excision repair and mismatch repair pathways in active DNA demethylation". en. In: *Nucleic Acids Res.* 44.8, pp. 3713–3727.

Guengerich, F Peter (1992). "Metabolic activation of carcinogens". In: *Pharmacology & therapeutics* 54.1, pp. 17–61.

Habraken, Y et al. (1996). "Binding of insertion/deletion DNA mismatches by the heterodimer of yeast mismatch repair proteins MSH2 and MSH3". en. In: *Curr. Biol.* 6.9, pp. 1185–1187.

Haiman, Christopher A et al. (2006). "Ethnic and racial differences in the smoking-related risk of lung cancer". In: *New England Journal of Medicine* 354.4, pp. 333–342.

Hanawalt, Philip C and Graciela Spivak (2008). "Transcription-coupled DNA repair: two decades of progress and surprises". In: *Nature reviews Molecular cell biology* 9.12, p. 958.

Hanford, Marsha G et al. (1998). "Microsatellite mutation rates in cancer cell lines deficient or proficient in mismatch repair". In: *Oncogene* 16.18, pp. 2389–2393.

Haradhvala, N J et al. (2018). "Distinct mutational signatures characterize concurrent loss of polymerase proofreading and mismatch repair". en. In: *Nat. Commun.* 9.1, p. 1746.

Haradhvala, Nicholas J et al. (2016). "Mutational strand asymmetries in cancer genomes reveal mechanisms of DNA damage and repair". In: *Cell* 164.3, pp. 538–549.

Harr, B, B Zangerl, and C Schlötterer (2000). "Removal of microsatellite interruptions by DNA replication slippage: phylogenetic evidence from Drosophila". en. In: *Mol. Biol. Evol.* 17.7, pp. 1001–1009.

Hartman, Phil S et al. (2014). "Ethyl methanesulfonate induces mutations in Caenorhabditis elegans embryos at a high frequency". In: *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* 766, pp. 44–48.

Hartman, Philip S et al. (1989). "Excision repair of UV radiation-induced DNA damage in Caenorhabditis elegans." In: *Genetics* 122.2, pp. 379–385.

Hashimoto, Satoru, Hirofumi Anai, and Katsuhiro Hanada (2016). "Mechanisms of interstrand DNA crosslink repair and human disorders". In: *Genes and Environment* 38.1, p. 9.

Hayward, Nicholas K et al. (2017). "Whole-genome landscapes of major melanoma subtypes". In: *Nature* 545.7653, p. 175.

Health, US Department of, Human Services, et al. (2016). *14th report on carcinogens (RoC)*.

Hecht, Jonathan L and George L Mutter (2006). "Molecular and pathologic aspects of endometrial carcinogenesis". en. In: *J. Clin. Oncol.* 24.29, pp. 4783–4791.

Hegi, Monika E et al. (2005). "MGMT gene silencing and benefit from temozolomide in glioblastoma". In: *New England Journal of Medicine* 352.10, pp. 997–1003.

Helleday, Thomas, Saeed Eshtad, and Serena Nik-Zainal (2014). "Mechanisms underlying mutational signatures in human cancers". en. In: *Nat. Rev. Genet.* 15.9, pp. 585–598.

Helleday, Thomas et al. (2008). "DNA repair pathways as targets for cancer therapy". In: *Nature Reviews Cancer* 8.3, p. 193.

Herman, J G et al. (1998). "Incidence and functional consequences of hMLH1 promoter hypermethylation in colorectal carcinoma". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 95.12, pp. 6870–6875.

Herman, JR et al. (1996). "UV-B increases (1979–1992) from decreases in total ozone". In: *Geophysical Research Letters* 23.16, pp. 2117–2120.

Hess, Marin T et al. (1997). "Base pair conformation-dependent excision of benzo [a] pyrene diol epoxide-guanine adducts by human nucleotide excision repair enzymes." In: *Molecular and cellular biology* 17.12, pp. 7069–7076.

Hillier, LaDeana W et al. (2005). "Genomics in C. elegans: so many genes, such a little worm". In: *Genome research* 15.12, pp. 1651–1660.

Hoang, Margaret L et al. (2016). "Genome-wide quantification of rare somatic mutations in normal human tissues using massively parallel sequencing". In: *Proceedings of the National Academy of Sciences* 113.35, pp. 9846–9851.

Hockemeyer, Dirk and Rudolf Jaenisch (2016). "Induced pluripotent stem cells meet genome editing". In: *Cell stem cell* 18.5, pp. 573–586.

Hoeijmakers, Jan HJ (2009). "DNA damage, aging, and cancer". In: *New England Journal of Medicine* 361.15, pp. 1475–1485.

Hollstein, M et al. (2017). "Base changes in tumour DNA have the power to reveal the causes and evolution of cancer". en. In: *Oncogene* 36.2, pp. 158–167.

Hollstein, Monica et al. (1991). "p53 mutations in human cancers". In: *Science* 253.5015, pp. 49–53.

Hope, Ian A (1999). *C. elegans: a practical approach*. Vol. 213. OUP Oxford.

Hosoya, Noriko and Kiyoshi Miyagawa (2014). "Targeting DNA damage response in cancer therapy". In: *Cancer science* 105.4, pp. 370–388.

Hsieh, Peggy and Kazuhiko Yamane (2008). "DNA mismatch repair: molecular mechanism, cancer, and ageing". In: *Mechanisms of ageing and development* 129.7-8, pp. 391–407.

Hu, Jinchuan et al. (2015). "Genome-wide analysis of human global and transcription-coupled excision repair of UV damage at single-nucleotide resolution". In: *Genes & development* 29.9, pp. 948–960.

Huang, Haimei (1981). "Ethyl Methanesulfonate (Ems) and Diethylnitrosamine (Den) Effect on Germ Cells of Drosophila Melanogaster." In:

Huang, Mi Ni et al. (2015). "MSIseq: Software for Assessing Microsatellite Instability from Catalogs of Somatic Mutations". en. In: *Sci. Rep.* 5, p. 13321.

Huang, Mi Ni et al. (2017a). "Genome-scale mutational signatures of aflatoxin in cells, mice, and human tumors". en. In: *Genome Res.* 27.9, pp. 1475–1486.

Huang, Tze-Yun et al. (2017b). "Parity-dependent hairpin configurations of repetitive DNA sequence promote slippage associated with DNA expansion". In: *Proceedings of the National Academy of Sciences* 114.36, pp. 9535–9540.

Huang, Xiaoqing, Damian Wojtowicz, and Teresa M Przytycka (2017). "Detecting presence of mutational signatures in cancer with confidence". In: *Bioinformatics* 34.2, pp. 330–337.

Huang, Yaling and Lei Li (2013). "DNA crosslinking damage and cancer-a tale of friend and foe". In: *Translational cancer research* 2.3, p. 144.

Ikehata, Hironobu and Tetsuya Ono (2011). "The mechanisms of UV mutagenesis". In: *Journal of radiation research* 52.2, pp. 115–125.

International, India Project Team of the et al. (2013). "Mutational landscape of gingivo-buccal oral squamous cell carcinoma reveals new recurrently-mutated genes and molecular subgroups". In: *Nature communications* 4, p. 2873.

Izsvák, Zsuzsanna, Yongming Wang, and Zoltán Ivics (2009). "Interactions of transposons with the cellular DNA repair machinery". In: *Transposons and the Dynamic Genome*. Springer, pp. 133–176.

Jackson, Stephen P (2002). "Sensing and repairing DNA double-strand breaks". In: *Carcinogenesis* 23.5, pp. 687–696.

Jackson, Stephen P and Jiri Bartek (2009). "The DNA-damage response in human biology and disease". In: *Nature* 461.7267, p. 1071.

Jager, Myrthe et al. (2019). "Deficiency of nucleotide excision repair is associated with mutational signature observed in cancer". In: *Genome Research*.

Johnson, Robert E et al. (2001). "Role of DNA polymerase η in the bypass of a (6-4) TT photoproduct". In: *Molecular and cellular biology* 21.10, pp. 3558–3563.

Jonnalagadda, Vidya S, Tetsuya Matsuguchi, and Bevin P Engelward (2005). "Interstrand crosslink-induced homologous recombination carries an increased risk of deletions and insertions". In: *Dna Repair* 4.5, pp. 594–605.

Kadyrov, Farid A et al. (2006). "Endonucleolytic function of MutLalpha in human mismatch repair". en. In: *Cell* 126.2, pp. 297–308.

Kaletta, Titus and Michael O Hengartner (2006). "Finding function in novel targets: C. elegans as a model organism". In: *Nature reviews Drug discovery* 5.5, p. 387.

Kalisperati, Polyxeni et al. (2017). "Inflammation, DNA damage, Helicobacter pylori and gastric tumorigenesis". In: *Frontiers in genetics* 8, p. 20.

Karran, Peter and Tomas Lindahl (1980). "Hypoxanthine in deoxyribonucleic acid: generation by heat-induced hydrolysis of adenine residues and release in free form by a deoxyribonucleic acid glycosylase from calf thymus". In: *Biochemistry* 19.26, pp. 6005–6011.

Kasar, S et al. (2015). "Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution". In: *Nature communications* 6, p. 8866.

Kato, Niyo et al. (2017). "Sensing and processing of DNA interstrand crosslinks by the mismatch repair pathway". In: *Cell reports* 21.5, pp. 1375–1385.

Kelderman, Sander, Ton N Schumacher, and Pia Kvistborg (2015). "Mismatch Repair-Deficient Cancers Are Targets for Anti-PD-1 Therapy". en. In: *Cancer Cell* 28.1, pp. 11–13.

Kew, Michael C (2013). "Aflatoxins as a cause of hepatocellular carcinoma." In: *Journal of Gastrointestinal & Liver Diseases* 22.3.

Kim, Hoon et al. (2015). "Whole-genome and multisector exome sequencing of primary and post-treatment glioblastoma reveals patterns of tumor evolution". en. In: *Genome Res.* 25.3, pp. 316–327.

Kim, Hyun Suk, Robert Hromas, and Suk-Hee Lee (2013). "Emerging features of dna double-strand break repair in humans". In: *New Research Directions in DNA Repair*. IntechOpen.

Kim, Jaegil et al. (2016). "Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors". en. In: *Nat. Genet.* 48.6, pp. 600–606.

Klug, William S, Michael R Cummings, et al. (2006). *Concepts of genetics*. Upper Saddle River, NJ: Pearson Education,

Knijnenburg, Theo A et al. (2018). "Genomic and Molecular Landscape of DNA Damage Repair Deficiency across The Cancer Genome Atlas". en. In: *Cell Rep.* 23.1, 239–254.e6.

Knisbacher, Binyamin A, Doron Gerber, and Erez Y Levanon (2016). "DNA editing by APOBECs: a genomic preserver and transformer". In: *Trends in Genetics* 32.1, pp. 16–28.

Knobel, Philip A and Thomas M Marti (2011). "Translesion DNA synthesis in the context of cancer research". en. In: *Cancer Cell Int.* 11, p. 39.

Koç, Ahmet et al. (2004). "Hydroxyurea arrests DNA replication by a mechanism that preserves basal dNTP pools". In: *Journal of Biological Chemistry* 279.1, pp. 223–230.

Kondo, Natsuko et al. (2010). "DNA damage induced by alkylating agents and repair pathways". In: *Journal of nucleic acids* 2010.

Koren, Amnon et al. (2012). "Differential relationship of DNA replication timing to different forms of human mutation and variation". In: *The American Journal of Human Genetics* 91.6, pp. 1033–1040.

Kramara, J, B Osia, and A Malkova (2018). "Break-induced replication: the where, the why, and the how". In: *Trends in Genetics* 34.7, pp. 518–531.

Krokan, Hans E and Magnar Bjørås (2013). "Base excision repair". In: *Cold Spring Harbor perspectives in biology* 5.4, a012583.

Krüger, Sandra and Rosario M Piro (2019). "decompTumor2Sig: identification of mutational signatures active in individual tumors". In: *BMC bioinformatics* 20.4, p. 152.

Kucab, Jill E et al. (2019). "A Compendium of Mutational Signatures of Environmental Agents". en. In: *Cell*.

Kunkel, Thomas A and Katarzyna Bebenek (2000). "DNA replication fidelity". In: *Annual review of biochemistry* 69.1, pp. 497–529.

Lachaud, Christophe et al. (2016). "Ubiquitinated Fancd2 recruits Fan1 to stalled replication forks to prevent genome instability". In: *Science* 351.6275, pp. 846–849.

Laghi, L, P Bianchi, and A Malesci (2008). "Differences and evolution of the methods for the assessment of microsatellite instability". en. In: *Oncogene* 27.49, pp. 6313–6321.

Lahtz, Christoph and Gerd P Pfeifer (2011). "Epigenetic changes of DNA repair genes in cancer". en. In: *J. Mol. Cell Biol.* 3.1, pp. 51–58.

Lang, Gregory I, Lance Parsons, and Alison E Gammie (2013). "Mutation rates, spectra, and genome-wide distribution of spontaneous mutations in mismatch repair deficient yeast". en. In: *G3* 3.9, pp. 1453–1465.

Lange, Sabine S, Kei-ichi Takata, and Richard D Wood (2011). "DNA polymerases and cancer". In: *Nature reviews cancer* 11.2, p. 96.

Lans, Hannes and Wim Vermeulen (2011). "Nucleotide Excision Repair inCaenorhabditis elegans". In: *Molecular Biology International* 2011, pp. 1–12.

Larsen, Nicolai B et al. (2017). "Stalled replication forks generate a distinct mutational signature in yeast". In: *Proceedings of the National Academy of Sciences* 114.36, pp. 9665–9670.

Laureti, Luisa et al. (2013). "Reduction of dNTP levels enhances DNA replication fidelity in vivo". In: *DNA repair* 12.4, pp. 300–305.

Lawes, D A, S SenGupta, and P B Boulos (2003). "The clinical importance and prognostic implications of microsatellite instability in sporadic cancer". en. In: *Eur. J. Surg. Oncol.* 29.3, pp. 201–212.

Le, Dung T et al. (2015). "PD-1 Blockade in Tumors with Mismatch-Repair Deficiency". en. In: *N. Engl. J. Med.* 372.26, pp. 2509–2520.

Le, Dung T et al. (2017). "Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade". In: *Science* 357.6349, pp. 409–413.

Lee, Daniel D and Sebastian H Seung (1999). "Learning the parts of objects by non-negative matrix factorization". In: *Nature* 401.6755, pp. 788–791.

Lee, Hojoon et al. (2015). "The Cancer Genome Atlas Clinical Explorer: a web and mobile interface for identifying clinical-genomic driver associations". en. In: *Genome Med.* 7, p. 112.

Lee, Hye Seung et al. (2002). "Distinct clinical features and outcomes of gastric cancers with microsatellite instability". en. In: *Mod. Pathol.* 15.6, pp. 632–640.

Lee, Raymond Y N et al. (2017). "WormBase 2017: molting into a new stage". In: *Nucleic acids research* 46.D1, pp. D869–D874.

Leinonen, Rasko et al. (2010). "The European nucleotide archive". In: *Nucleic acids research* 39.suppl_1, pp. D28–D31.

Lelieveld, Stefan H et al. (2015). "Comparison of exome and genome sequencing technologies for the complete capture of protein-coding regions". In: *Human mutation* 36.8, pp. 815–822.

Leung, Maxwell CK et al. (2010). "Caenorhabditis elegans generates biologically relevant levels of genotoxic metabolites from aflatoxin B1 but not benzo [a] pyrene in vivo". In: *Toxicological Sciences* 118.2, pp. 444–453.

Li, Heng and Richard Durbin (2009). "Fast and accurate short read alignment with Burrows–Wheeler transform". In: *bioinformatics* 25.14, pp. 1754–1760.

Li, Wentao et al. (2017a). "Human genome-wide repair map of DNA damage caused by the cigarette smoke carcinogen benzo [a] pyrene". In: *Proceedings of the National Academy of Sciences* 114.26, pp. 6752–6757.

Li, XC et al. (2018). "A mutational signature associated with alcohol consumption and prognostically significantly mutated driver genes in esophageal squamous cell carcinoma". In: *Annals of Oncology* 29.4, pp. 938–944.

Li, Xiangchun et al. (2016). "Distinct subtypes of gastric cancer defined by molecular characterization include novel mutational signatures with prognostic capability". In: *Cancer research* 76.7, pp. 1724–1732.

Li, Xuan and Wolf-Dietrich Heyer (2008). "Homologous recombination in DNA repair and DNA damage tolerance". In: *Cell research* 18.1, p. 99.

Li, Y et al. (2017b). "Patterns of structural variation in human cancer". In: *bioRxiv*.

Li, YWAN FENG, Sang-Tae Kim, and Aziz Sancar (1993). "Evidence for lack of DNA photoreactivating enzyme in humans." In: *Proceedings of the National Academy of Sciences* 90.10, pp. 4389–4393.

Lieber, Michael R (2010). "The mechanism of double-strand DNA break repair by the nonhomologous DNA end-joining pathway". In: *Annual review of biochemistry* 79, pp. 181–211.

Lin, Ying-Chih et al. (2014). "Molecular basis of aflatoxin-induced mutagenesis—role of the aflatoxin B1-formamidopyrimidine adduct". In: *Carcinogenesis* 35.7, pp. 1461–1468.

Lindahl, Tomas and DE Barnes (2000). "Repair of endogenous DNA damage". In: *Cold Spring Harbor symposia on quantitative biology*. Vol. 65. Cold Spring Harbor Laboratory Press, pp. 127–134.

Lindahl, Tomas et al. (1993). "Instability and decay of the primary structure of DNA". In: *nature* 362.6422, pp. 709–715.

Lisby, Michael and Rodney Rothstein (2015). "Cell biology of mitotic recombination". In: *Cold Spring Harbor perspectives in biology* 7.3, a016535.

Little, John B (1993). "Cellular, molecular, and carcinogenic effects of radiation". In: *Hematology/Oncology Clinics* 7.2, pp. 337–352.

Liu, David et al. (2017a). "Mutational patterns in chemotherapy resistant muscle-invasive bladder cancer". In: *Nature communications* 8.1, p. 2193.

Liu, Gang et al. (2017b). "Genomics alterations of metastatic and primary tissues across 15 cancer types". In: *Scientific reports* 7.1, p. 13262.

Liu, Leroy F and James C Wang (1987). "Supercoiling of the DNA template during transcription". In: *Proceedings of the National Academy of Sciences* 84.20, pp. 7024–7027.

Liu, Qian et al. (2019). "NanoMod: a computational tool to detect DNA modifications using Nanopore long-read sequencing data". In: *BMC genomics* 20.1, p. 78.

Liu, Yunhua et al. (2015). "TP53 loss creates therapeutic vulnerability in colorectal cancer". en. In: *Nature* 520.7549, pp. 697–701.

Lobitz, Stephan and Eunike Velleuer (2006). "Guido Fanconi (1892–1979): a jack of all trades". In: *Nature Reviews Cancer* 6.11, p. 893.

Lomax, ME, LK Folkes, and P O'Neill (2013). "Biological consequences of radiation-induced DNA damage: relevance to radiotherapy". In: *Clinical oncology* 25.10, pp. 578–585.

Luftig, Micah A (2014). "Viruses and the DNA damage response: activation and antagonism". In: *Annual Review of Virology* 1, pp. 605–625.

Lujan, Scott A, Alan B Clark, and Thomas A Kunkel (2015). "Differences in genome-wide repeat sequence instability conferred by proofreading and mismatch repair defects". en. In: *Nucleic Acids Res.* 43.8, pp. 4067–4074.

Lujan, Scott A et al. (2012). "Mismatch repair balances leading and lagging strand DNA replication fidelity". en. In: *PLoS Genet.* 8.10, e1003016.

Lujan, Scott A et al. (2014). "Heterogeneous polymerase fidelity and mismatch repair bias genome variation and composition". en. In: *Genome Res.* 24.11, pp. 1751–1764.

Ma, Jennifer et al. (2018). "The therapeutic significance of mutational signatures from DNA repair deficiency in cancer". In: *Nature communications* 9.1, pp. 1–12.

Maaten, Laurens van der and Geoffrey Hinton (2008). "Visualizing Data using t-SNE". In: *J. Mach. Learn. Res.* 9.Nov, pp. 2579–2605.

Macintyre, Geoff et al. (2018). "Copy number signatures and mutational processes in ovarian carcinoma". In: *Nature genetics* 50.9, p. 1262.

Maharjan, Ram and Thomas Ferenci (2014). "Mutational signatures indicative of environmental stress in bacteria". In: *Molecular biology and evolution* 32.2, pp. 380–391.

Mansour, Wael Y, Tim Rhein, and Jochen Dahm-Daphi (2010). "The alternative end-joining pathway for repair of DNA double-strand breaks requires PARP1 but is not dependent upon microhomologies". In: *Nucleic acids research* 38.18, pp. 6065–6077.

Marsico, Giovanni et al. (2019). "Whole genome experimental maps of DNA G-quadruplexes in multiple species". In: *Nucleic acids research* 47.8, pp. 3862–3874.

Martincorena, Iñigo et al. (2015). "High burden and pervasive positive selection of somatic mutations in normal human skin". In: *Science* 348.6237, pp. 880–886.

Martincorena, Iñigo et al. (2017). "Universal Patterns of Selection in Cancer and Somatic Tissues". en. In: *Cell* 171.5, 1029–1041.e21.

Martincorena, Iñigo et al. (2018). "Somatic mutant clones colonize the human esophagus with age". In: *Science* 362.6417, pp. 911–917.

Masuda, Keiji et al. (2009). "A critical role for REV1 in regulating the induction of C: G transitions and A: T mutations during Ig gene hypermutation". In: *The Journal of Immunology* 183.3, pp. 1846–1850.

Masutani, Chikahide et al. (1999). "The XPV (xeroderma pigmentosum variant) gene encodes human DNA polymerase η". In: *Nature* 399.6737, p. 700.

Matsuda, Toshiro et al. (2001). "Error rate and specificity of human and murine DNA polymerase η". In: *Journal of molecular biology* 312.2, pp. 335–346.

Matsumura, Shoji et al. (2018). "A genome-wide mutation analysis method enabling high-throughput identification of chemical mutagen signatures". In: *Scientific Reports* 8.1.

Mayles, WPM et al. (2010). "Survey of the availability and use of advanced radiotherapy technology in the UK". In: *Clinical Oncology* 22.8, pp. 636–642.

McCreery, Melissa Q et al. (2015). "Evolution of metastasis revealed by mutational landscapes of chemically induced skin cancers". In: *Nature medicine* 21.12, p. 1514.

McGregor, W Glemm et al. (1991). "Cell cycle-dependent strand bias for UV-induced mutations in the transcribed strand of excision repair-proficient human fibroblasts but not in repair-deficient cells." In: *Molecular and cellular biology* 11.4, pp. 1927–1934.

McLaren, William et al. (2016). "The Ensembl Variant Effect Predictor". en. In: *Genome Biol.* 17.1, p. 122.

McMahill, Melissa S, Caroline W Sham, and Douglas K Bishop (2007). "Synthesis-dependent strand annealing in meiosis". In: *PLoS biology* 5.11, e299.

Meier, Bettina and Anton Gartner (2014). "Having a direct look: analysis of DNA damage and repair mechanisms by next generation sequencing". en. In: *Exp. Cell Res.* 329.1, pp. 35–41.

Meier, Bettina et al. (2014). "C. elegans whole-genome sequencing reveals mutational signatures related to carcinogens and DNA repair deficiency". en. In: *Genome Res.* 24.10, pp. 1624–1636.

Meier, Bettina et al. (2018). "Mutational signatures of DNA mismatch repair deficiency in C. elegans and human cancers". en. In: *Genome Res.* 28.5, pp. 666–675.

Mimaki, Sachiyo et al. (2016). “Hypermutation and unique mutational signatures of occupational cholangiocarcinoma in printing workers exposed to haloalkanes”. In: *Carcinogenesis* 37.8, pp. 817–826.

Minca, Eugen C and David Kowalski (2010). “Replication fork stalling by bulky DNA damage: localization at active origins and checkpoint modulation”. In: *Nucleic acids research* 39.7, pp. 2610–2623.

Mishra, Sweta and Johnathan R Whetstine (2016). “Different facets of copy number changes: permanent, transient, and adaptive”. In: *Molecular and cellular biology* 36.7, pp. 1050–1063.

Miyaki, Michiko et al. (1997). “Germline mutation of MSH6 as the cause of hereditary nonpolyposis colorectal cancer”. In: *Nature Genetics* 17.3, pp. 271–272.

Moore, Harold W and Richard Czerniak (1981). “Naturally occurring quinones as potential bioreductive alkylating agents”. In: *Medicinal research reviews* 1.3, pp. 249–280.

Morganella, Sandro et al. (2016). “The topography of mutational processes in breast cancer genomes”. en. In: *Nat. Commun.* 7, p. 11383.

Mosteller, Frederick and John W Tukey (1968). “Data analysis, including statistics”. In: *Handbook of social psychology* 2, pp. 80–203.

Neal, Radford M and Others (2011). “MCMC using Hamiltonian dynamics”. In: *Handbook of Markov Chain Monte Carlo* 2.11, p. 2.

Ng, Alvin W T et al. (2017). “Aristolochic acids and their derivatives are widely implicated in liver cancers in Taiwan and throughout Asia”. en. In: *Sci. Transl. Med.* 9.412.

Nicolaides, N C et al. (1994). “Mutations of two PMS homologues in hereditary nonpolyposis colon cancer”. en. In: *Nature* 371.6492, pp. 75–80.

Nik-Zainal, Serena et al. (2012a). “Mutational processes molding the genomes of 21 breast cancers”. en. In: *Cell* 149.5, pp. 979–993.

Nik-Zainal, Serena et al. (2012b). “The life history of 21 breast cancers”. In: *Cell* 149.5, pp. 994–1007.

Nik-Zainal, Serena et al. (2016). “Landscape of somatic mutations in 560 breast cancer whole-genome sequences”. en. In: *Nature* 534.7605, pp. 47–54.

Nitiss, John L (2009). “Targeting DNA topoisomerase II in cancer chemotherapy”. In: *Nature Reviews Cancer* 9.5, p. 338.

Niu, Beifang et al. (2014). “MSIsensor: microsatellite instability detection using paired tumor-normal sequence data”. en. In: *Bioinformatics* 30.7, pp. 1015–1016.

Noll, David M, Tracey McGregor Mason, and Paul S Miller (2006). “Formation and repair of interstrand cross-links in DNA”. In: *Chemical reviews* 106.2, pp. 277–301.

Nouspikel, TDNA (2009). “DNA repair in mammalian cells”. In: *Cellular and Molecular Life Sciences* 66.6, pp. 994–1009.

Nowak, Jonathan A et al. (2017). “Detection of Mismatch Repair Deficiency and Microsatellite Instability in Colorectal Adenocarcinoma by Targeted Next-Generation Sequencing”. en. In: *J. Mol. Diagn.* 19.1, pp. 84–91.

O’Donovan, Anne et al. (1994). “XPG endonuclease makes the 3 incision in human DNA nucleotide excision repair”. In: *Nature* 371.6496, p. 432.

Olivier, Magali et al. (2014). “Modelling mutational landscapes of human cancers in vitro”. In: *Scientific reports* 4, p. 4482.

Omichessan, Hanane, Gianluca Severi, and Vittorio Perduca (2019). “Computational tools to detect signatures of mutational processes in DNA from tumours: a review and empirical comparison of performance”. In: *bioRxiv*, p. 483982.

Palles, Claire et al. (2013). “Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas”. en. In: *Nat. Genet.* 45.2, pp. 136–144.

Papadopoulos, N et al. (1994). “Mutation of a mutL homolog in hereditary colon cancer”. In: *Science* 263.5153, pp. 1625–1629.

Pearl, Laurence H et al. (2015). “Therapeutic opportunities within the DNA damage response”. en. In: *Nat. Rev. Cancer* 15.3, pp. 166–180.

Petljak, Mia et al. (2019). “Characterizing Mutational Signatures in Human Cancer Cell Lines Reveals Episodic APOBEC Mutagenesis”. en. In: *Cell* 176.6, 1282–1294.e20.

Petruseva, IO, AN Evdokimov, and OI Lavrik (2014). “Molecular mechanism of global genome nucleotide excision repair”. In: *Acta Naturae ( )* 6.1 (20).

Pettersen, Henrik Sahlin et al. (2015). “AID expression in B-cell lymphomas causes accumulation of genomic uracil and a distinct AID mutational signature”. In: *DNA repair* 25, pp. 60–71.

Pfau, Wolfgang, Heinz H Schmeiser, and Manfred Wiessler (1990). “Aristolochic acid binds covalently to the exocyclic amino group of purine nucleotides in DNA”. In: *Carcinogenesis* 11.2, pp. 313–319.

Pfeifer, Gerd P (2010). “Environmental exposures and mutational patterns of cancer genomes”. In: *Genome medicine* 2.8, p. 54.

Phillips, David H (2002). “Smoking-related DNA and protein adducts in human tissues”. In: *Carcinogenesis* 23.12, pp. 1979–2004.

Pilati, Camilla et al. (2017). “Mutational signature analysis identifies MUTYH deficiency in colorectal cancers and adrenocortical carcinomas”. In: *The Journal of pathology* 242.1, pp. 10–15.

Piovesan, Allison et al. (2019). "On the length, weight and GC content of the human genome". In: *BMC research notes* 12.1, p. 106.

Pleasance, Erin D et al. (2010). "A comprehensive catalogue of somatic mutations from a human cancer genome". In: *Nature* 463.7278, p. 191.

Pluciennik, Anna et al. (2010). "PCNA function in the activation and strand direction of MutLα endonuclease in mismatch repair". In: *Proceedings of the National Academy of Sciences* 107.37, pp. 16066–16071.

Poetsch, Anna R, Simon J Boulton, and Nicholas M Luscombe (2018). "Genomic landscape of oxidative DNA damage and repair reveals regioselective protection from mutagenesis". In: *Genome biology* 19.1, p. 215.

Polak, Paz et al. (2017). "A mutational signature reveals alterations underlying deficient homologous recombination repair in breast cancer". In: *Nature genetics* 49.10, p. 1476.

Poon, Song Ling et al. (2013). "Genome-wide mutational signatures of aristolochic acid and its application as a screening tool". en. In: *Sci. Transl. Med.* 5.197, 197ra101.

Poon, Song Ling et al. (2014). "Mutation signatures of carcinogen exposure: genome-wide detection and new opportunities for cancer prevention". en. In: *Genome Med.* 6.3, p. 24.

Poon, Song Ling et al. (2015). "Mutation signatures implicate aristolochic acid in bladder cancer development". en. In: *Genome Med.* 7.1, p. 38.

Pourkarimi, Ehsan, James M Bellush, and Iestyn Whitehouse (2016). "Spatiotemporal coupling and decoupling of gene transcription with DNA replication origins during embryogenesis in C. elegans". In: *Elife* 5, e21728.

Povirk, Lawrence F and David E Shuker (1994). "DNA damage and mutagenesis induced by nitrogen mustards". In: *Mutation research/reviews in genetic toxicology* 318.3, pp. 205–226.

Priestley, Peter et al. (2018). "Pan-cancer whole genome analyses of metastatic solid tumors". In: *bioRxiv*, p. 415133.

Prioleau, Marie-Noëlle and David M MacAlpine (2016). "DNA replication origins—where do we begin?" In: *Genes & development* 30.15, pp. 1683–1697.

Puente, Xose S et al. (2011). "Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia". In: *Nature* 475.7354, p. 101.

Rabik, Cara A and M Eileen Dolan (2007). "Molecular mechanisms of resistance and toxicity associated with platinating agents". In: *Cancer treatment reviews* 33.1, pp. 9–23.

Ramazzotti, Daniele et al. (2018). "De novo mutational signature discovery in tumor genomes using SparseSignatures". In: *bioRxiv*, p. 384834.

Rastogi, Rajesh P et al. (2010). "Molecular mechanisms of ultraviolet radiation-induced DNA damage and repair". In: *Journal of nucleic acids* 2010.

Rausch, Tobias et al. (2012). "DELLY: structural variant discovery by integrated paired-end and split-read analysis". en. In: *Bioinformatics* 28.18, pp. i333–i339.

Rechkoblit, Olga et al. (2002). "trans-Lesion synthesis past bulky benzo [a] pyrene diol epoxide N 2-dG and N 6-dA lesions catalyzed by DNA bypass polymerases". In: *Journal of Biological Chemistry* 277.34, pp. 30488–30494.

Reijns, Martin AM et al. (2015). "Lagging-strand replication shapes the mutational landscape of the genome". In: *Nature* 518.7540, p. 502.

Riaz, Nadeem et al. (2017). "Pan-cancer analysis of bi-allelic alterations in homologous recombination DNA repair genes". In: *Nature communications* 8.1, p. 857.

Rippey, JCR and MI Stallwood (2005). "Nine cases of accidental exposure to dimethyl sulphate—a potential chemical weapon". In: *Emergency medicine journal* 22.12, pp. 878–879.

Rivas, Miguel et al. (2011). "Ultraviolet light exposure influences skin cancer in association with latitude". In: *Oncology reports* 25.4, pp. 1153–1159.

Rivlin, Noa et al. (2011). "Mutations in the p53 Tumor Suppressor Gene: Important Milestones at the Various Steps of Tumorigenesis". en. In: *Genes Cancer* 2.4, pp. 466–474.

Roberts, Steven A and Dmitry A Gordenin (2014). "Hypermutation in human cancer genomes: footprints and mechanisms". en. In: *Nat. Rev. Cancer* 14.12, pp. 786–800.

Roberts, Steven A et al. (2012). "Clustered mutations in yeast and in human cancers can arise from damaged long single-strand DNA regions". In: *Molecular cell* 46.4, pp. 424–435.

Roberts, Steven A et al. (2013). "An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers". en. In: *Nat. Genet.* 45.9, pp. 970–976.

Robertson, AB et al. (2009). "DNA repair in mammalian cells". In: *Cellular and molecular life sciences* 66.6, pp. 981–993.

Robinson, Welles, Roded Sharan, and Mark DM Leiserson (2019). "Modeling clinical and molecular covariates of mutational process activity in cancer". In: *Bioinformatics* 35.14, pp. i492–i500.

Rodgers, Kasey and Mitch McVey (2016). "Error-prone repair of DNA double-strand breaks". In: *Journal of cellular physiology* 231.1, pp. 15–24.

Roerink, Sophie F, Robin van Schendel, and Marcel Tijsterman (2014). "Polymerase theta-mediated end joining of replication-associated DNA breaks in C. elegans". In: *Genome Res.* 24.6, pp. 954–962.

Rosales, Rafael A et al. (2016). "signeR: an empirical Bayesian approach to mutational signature discovery". In: *Bioinformatics* 33.1, pp. 8–16.

Rosenthal, Rachel et al. (2016). "DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution". en. In: *Genome Biol.* 17, p. 31.

Russell, WL and EM Kelly (1982). "Mutation frequencies in male mice and the estimation of genetic hazards of radiation in men". In: *Proceedings of the National Academy of Sciences* 79.2, pp. 542–544.

Sale, Julian E (2013). "Translesion DNA synthesis and mutagenesis in eukaryotes". In: *Cold Spring Harbor perspectives in biology* 5.3, a012708.

Sale, Julian E, Alan R Lehmann, and Roger Woodgate (2012). "Y-family DNA polymerases and their role in tolerance of cellular DNA damage". In: *Nature reviews Molecular cell biology* 13.3, p. 141.

Sancar, Aziz et al. (2004). "Molecular mechanisms of mammalian DNA repair and the DNA damage checkpoints". In: *Annual review of biochemistry* 73.1, pp. 39–85.

Sanders, Mathijs A et al. (2018). "MBD4 guards against methylation damage and germ line deficiency predisposes to clonal hematopoiesis and early-onset AML". In: *Blood* 132.14, pp. 1526–1534.

Sankar, T Sabari et al. (2016). "The nature of mutations induced by replication–transcription collisions". In: *Nature* 535.7610, p. 178.

Schendel, Robin van et al. (2016). "Genomic scars generated by polymerase theta reveal the versatile mechanism of alternative end-joining". In: *PLoS genetics* 12.10, e1006368.

Schiller, John T and Douglas R Lowy (2010). "Vaccines to prevent infections by oncoviruses". In: *Annual review of microbiology* 64, pp. 23–41.

Schimmel, Joost et al. (2017). "Mutational signatures of non-homologous and polymerase theta-mediated end-joining in embryonic stem cells". In: *The EMBO journal* 36.24, pp. 3634–3649.

Schumacher, April J, Dwight V Nissley, and Reuben S Harris (2005). "APOBEC3G hypermutates genomic DNA and inhibits Ty1 retrotransposition in yeast". In: *Proceedings of the National Academy of Sciences* 102.28, pp. 9854–9859.

Secrier, Maria et al. (2016). "Mutational signatures in esophageal adenocarcinoma define etiologically distinct subgroups with therapeutic relevance". In: *Nature genetics* 48.10, p. 1131.

Segovia, Romulo, Annie S Tam, and Peter C Stirling (2015). "Dissecting genetic and environmental mutation signatures with model organisms". en. In: *Trends Genet.* 31.8, pp. 465–474.

Seki, Mineaki, Federica Marini, and Richard D Wood (2003). "POLQ (Pol θ), a DNA polymerase and DNA-dependent ATPase in human cells". In: *Nucleic acids research* 31.21, pp. 6117–6126.

Seol, Ja-Hwan, Eun Yong Shim, and Sang Eun Lee (2018). "Microhomology-mediated end joining: Good, bad and ugly". In: *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* 809, pp. 81–87.

Seplyarskiy, Vladimir B et al. (2016). "APOBEC-induced mutations in human cancers are strongly enriched on the lagging DNA strand during replication". In: *Genome research* 26.2, pp. 174–182.

Sequencing Consortium*, The C. elegans (1998). "Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology". In: *Science* 282.5396, pp. 2012–2018. ISSN: 0036-8075. DOI: 10.1126/science.282.5396.2012. eprint: https://science.sciencemag.org/content/282/5396/2012.full.pdf. URL: https://science.sciencemag.org/content/282/5396/2012.

Shah, DJ, RK Sachs, and DJ Wilson (2012). "Radiation-induced cancer: a modern view". In: *The British journal of radiology* 85.1020, e1166–e1173.

Shaheen, Montaser et al. (2011). "Synthetic lethality: exploiting the addiction of cancer to DNA repair". In: *Blood* 117.23, pp. 6074–6082.

Shaye, Daniel D and Iva Greenwald (2011). "OrthoList: a compendium of C. elegans genes with human orthologs". In: *PloS one* 6.5, e20085.

Sherry, Stephen T et al. (2001). "dbSNP: the NCBI database of genetic variation". In: *Nucleic acids research* 29.1, pp. 308–311.

Shinbrot, Eve et al. (2014). "Exonuclease mutations in DNA polymerase epsilon reveal replication strand specific mutation patterns and human origins of replication". en. In: *Genome Res.* 24.11, pp. 1740–1750.

Shiraishi, Yuichi et al. (2015). "A simple model-based approach to inferring and visualizing cancer mutation signatures". In: *PLoS genetics* 11.12, e1005657.

Shlien, Adam et al. (2015). "Combined hereditary and somatic mutations of replication error repair genes result in rapid onset of ultra-hypermutated cancers". en. In: *Nat. Genet.* 47.3, pp. 257–262.

Sidorenko, Victoria S et al. (2012). "Lack of recognition by global-genome nucleotide excision repair accounts for the high mutagenicity and persistence of aristolactam-DNA adducts". en. In: *Nucleic Acids Res.* 40.6, pp. 2494–2505.

Simpson, AndrewJ G (1997). "The natural somatic mutation frequency and human carcinogenesis". In: *Advances in cancer research*. Vol. 71. Elsevier, pp. 209–240.

Strand, M et al. (1993). "Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair". en. In: *Nature* 365.6443, pp. 274–276.

Stratton, Michael R, Peter J Campbell, and Andrew P Futreal (2009). "The cancer genome". In: *Nature* 458.7239, pp. 719–724.

Supek, Fran and Ben Lehner (2015). "Differential DNA mismatch repair underlies mutation rate variation across the human genome". en. In: *Nature* 521.7550, pp. 81–84.

— (2017). "Clustered mutation signatures reveal that error-prone DNA repair targets mutations to active genes". In: *Cell* 170.3, pp. 534–547.

Swenberg, James A et al. (2010). "Endogenous versus exogenous DNA adducts: their role in carcinogenesis, epidemiology, and risk assessment". In: *Toxicological sciences* 120.suppl_1, S130–S145.

Sykora, Peter et al. (2018). "Next generation high throughput DNA damage detection platform for genotoxic compound screening". In: *Scientific reports* 8.1, p. 2771.

Szikriszt, Bernadett et al. (2016). "A comprehensive survey of the mutagenic impact of common cancer cytotoxics". en. In: *Genome Biol.* 17, p. 99.

Taira, Kentaro et al. (2013). "Distinct pathways for repairing mutagenic lesions induced by methylating and ethylating agents". en. In: *Mutagenesis* 28.3, pp. 341–350.

Tam, Annie S, Jeffrey SC Chu, and Ann M Rose (2016). "Genome-wide mutational signature of the chemotherapeutic agent mitomycin C in Caenorhabditis elegans". In: *G3: Genes, Genomes, Genetics* 6.1, pp. 133–140.

Tasaki, Eisuke et al. (2018). "High expression of the breast cancer susceptibility gene BRCA1 in long-lived termite kings". In: *Aging (Albany NY)* 10.10, p. 2668.

Taylor, Benjamin Jm et al. (2013). "DNA deaminases induce break-associated mutation showers with implication of APOBEC3B and 3A in breast cancer kataegis". en. In: *Elife* 2, e00534.

Taylor, Martin RG et al. (2015). "Rad51 paralogs remodel pre-synaptic Rad51 filaments to stimulate homologous recombination". In: *Cell* 162.2, pp. 271–286.

The 1000 Genomes Project Consortium (2015). "A global reference for human genetic variation". In: *Nature* 526.7571, pp. 68–74.

Thompson, Owen et al. (2013). "The million mutation project: a new approach to genetics in Caenorhabditis elegans". In: *Genome research* 23.10, pp. 1749–1762.

Tian-Shung, Wu et al. (2005). "Chemical constituents and pharmacology of Aristolochi species". In: *Studies in Natural Products Chemistry*. Vol. 32. Elsevier, pp. 855–1018.

Tijsterman, Marcel, Joris Pothof, and Ronald H A Plasterk (2002). "Frequent germline mutations and somatic repeat instability in DNA mismatch-repair-deficient Caenorhabditis elegans". en. In: *Genetics* 161.2, pp. 651–660.

Tomasetti, Cristian, Bert Vogelstein, and Giovanni Parmigiani (2013). "Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation". In: *Proc. Natl. Acad. Sci. U. S. A.* 110.6, pp. 1999–2004.

Tomasetti, Cristian et al. (2015). "Only three driver gene mutations are required for the development of lung and colorectal cancers". In: *Proc. Natl. Acad. Sci. U. S. A.* 112.1, pp. 118–123.

Tomkova, Marketa et al. (2018). "Mutational signature distribution varies with DNA replication timing and strand asymmetry". In: *Genome biology* 19.1, p. 129.

Tricarico, Rossella et al. (2015a). "Involvement of MBD4 inactivation in mismatch repair-deficient tumorigenesis". In: *Oncotarget* 6.40, p. 42892.

Tricarico, Rossella et al. (2015b). "Involvement of MBD4 inactivation in mismatch repair-deficient tumorigenesis". en. In: *Oncotarget* 6.40, pp. 42892–42904.

Trucco, Lucas D et al. (2019). "Ultraviolet radiation–induced DNA damage is prognostic for outcome in melanoma". In: *Nature medicine* 25.2, p. 221.

Tung, Emily WY et al. (2014). "Benzo [a] pyrene increases DNA double strand break repair in vitro and in vivo: a possible mechanism for benzo [a] pyrene-induced toxicity". In: *Mutation Research/Genetic Toxicology and Environmental Mutagenesis* 760, pp. 64–69.

Uringa, Evert-Jan et al. (2010). "RTEL1: an essential helicase for telomere maintenance and the regulation of homologous recombination". In: *Nucleic acids research* 39.5, pp. 1647–1655.

Van Hoeck, Arne et al. (2019). "Portrait of a cancer: mutational signature analyses for cancer diagnostics". In: *BMC cancer* 19.1, p. 457.

Van Loo, Peter et al. (2010). "Allele-specific copy number analysis of tumors". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 107.39, pp. 16910–16915.

Van Schendel, Robin et al. (2015). "Polymerase Θ is a key driver of genome evolution and of CRISPR/Cas9-mediated mutagenesis". In: *Nature communications* 6, p. 7394.

Viel, Alessandra et al. (2017). "A specific mutational signature associated with DNA 8-oxoguanine persistence in MUTYH-defective colorectal cancer". In: *EBioMedicine* 20, pp. 39–49.

Vineis, Paolo and Christopher P Wild (2014). "Global cancer patterns: causes and prevention". In: *The Lancet* 383.9916, pp. 549–557.

Viterbi, Andrew (1967). "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm". In: *IEEE transactions on Information Theory* 13.2, pp. 260–269.

Vogelstein, Bert and Kenneth W Kinzler (2015). "The Path to Cancer –Three Strikes and You're Out". In: *N. Engl. J. Med.* 373.20, pp. 1895–1898.

Vogelstein, Bert, Drew M Pardoll, and Donald S Coffey (1980). "Supercoiled loops and eucaryotic DNA replication". In: *Cell* 22.1, pp. 79–85.

Volkova, Nadezda V et al. (2019). "Mutational signatures are jointly shaped by DNA damage and repair". In: *bioRxiv*, p. 686295.

Wakelin, Laurence PG (1986). "Polyfunctional DNA intercalating agents". In: *Medicinal research reviews* 6.3, pp. 275–340.

Wang, Man-Tzu et al. (2015). "K-Ras Promotes Tumorigenicity through Suppression of Non-canonical Wnt Signaling". en. In: *Cell* 163.5, pp. 1237–1251.

Wang, Victor G, Hyunsoo Kim, and Jeffrey H Chuang (2018). "Whole-exome sequencing capture kit biases yield false negative mutation calls in TCGA cohorts". en. In: *PLoS One* 13.10, e0204912.

Ward, Elizabeth et al. (2004). "Cancer disparities by race/ethnicity and socioeconomic status". In: *CA: a cancer journal for clinicians* 54.2, pp. 78–93.

Ward, JF (1994). "The complexity of DNA damage: relevance to biological consequences". In: *International journal of radiation biology* 66.5, pp. 427–432.

Watson, William AF (1964). "Evidence of an essential difference between the genetical effects of mono-and bi-functional alkylating agents". In: *Zeitschrift für Vererbungslehre* 95.4, pp. 374–378.

Weghorn, Donate and Shamil Sunyaev (2017). "Bayesian inference of negative and positive selection in human cancers". en. In: *Nat. Genet.* 49.12, pp. 1785–1788.

Weinstein, John N et al. (2013). "The cancer genome atlas pan-cancer analysis project". In: *Nature genetics* 45.10, p. 1113.

Westcott, Peter MK et al. (2015). "The mutational landscapes of genetic and chemical models of Kras-driven lung cancer". In: *Nature* 517.7535, p. 489.

Wheless, Lee et al. (2012). "A community-based study of nucleotide excision repair polymorphisms in relation to the risk of non-melanoma skin cancer". en. In: *J. Invest. Dermatol.* 132.5, pp. 1354–1362.

Wijen, John PH, Madeleine JM Nivard, and Ekkehart W Vogel (2000). "The in vivo genetic activity profile of the monofunctional nitrogen mustard 2-chloroethylamine differs drastically from its bifunctional counterpart mechlorethamine". In: *Carcinogenesis* 21.10, pp. 1859–1867.

Willers, H, J Dahm-Daphi, and SN Powell (2004). "Repair of radiation damage to DNA". In: *British Journal of Cancer* 90.7, p. 1297.

Wilson III, David M and Vilhelm A Bohr (2007). "The mechanics of base excision repair, and its relationship to aging and disease". In: *DNA repair* 6.4, pp. 544–559.

Wojtowicz, Damian et al. (2019). "Hidden Markov models lead to higher resolution maps of mutation signature activity in cancer". In: *Genome medicine* 11.1, p. 49.

Wyatt, Michael D and Douglas L Pittman (2006). "Methylating agents and DNA repair responses: Methylated bases and sources of strand breaks". In: *Chemical research in toxicology* 19.12, pp. 1580–1594.

Xu, Hanxiao et al. (2018). "Organoid technology and applications in cancer research". In: *Journal of hematology & oncology* 11.1, pp. 1–15.

Xue, Weiling and David Warshawsky (2005). "Metabolic activation of polycyclic and heterocyclic aromatic hydrocarbons and DNA damage: a review". In: *Toxicology and applied pharmacology* 206.1, pp. 73–93.

Yadav, Vinod Kumar, James DeGregori, and Subhajyoti De (2016). "The landscape of somatic mutations in protein coding genes in apparently benign human tissues carries signatures of relaxed purifying selection". In: *Nucleic acids research* 44.5, pp. 2075–2084.

Yang, Y et al. (1999). "Analysis of yeast pms1, msh2, and mlh1 mutators points to differences in mismatch correction efficiencies between prokaryotic and eukaryotic cells". en. In: *Mol. Gen. Genet.* 261.4-5, pp. 777–787.

Yao, X et al. (1999). "Different mutator phenotypes in Mlh1- versus Pms2-deficient mice". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 96.12, pp. 6850–6855.

Ye, Kai et al. (2009). "Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads". en. In: *Bioinformatics* 25.21, pp. 2865–2871.

Yoon, Jung-Hoon, Louise Prakash, and Satya Prakash (2010). "Error-free replicative bypass of (6–4) photoproducts by DNA polymerase ζ in mouse and human cells". In: *Genes & development* 24.2, pp. 123–128.

Yoon, Jung-Hoon et al. (2017). "Translesion synthesis DNA polymerases promote error-free replication through the minor-groove DNA adduct 3-deaza-3-methyladenine". en. In: *J. Biol. Chem.* 292.45, pp. 18682–18688.

Yoon, Jung-Hoon et al. (2019). "Error-Prone Replication through UV Lesions by DNA Polymerase ϑ Protects against Skin Cancers". In: *Cell* 176.6, 1295–1309.e15.

Zapatka, Marc et al. (2018). "The landscape of viral associations in human cancers". In: *bioRxiv*, p. 465757.

Zeng, Xianmin et al. (2001). "DNA polymerase η is an AT mutator in somatic hypermutation of immunoglobulin variable genes". In: *Nature immunology* 2.6, p. 537.

Žgur-Bertok, Darja (2013). "DNA damage repair and bacterial pathogens". In: *PLoS pathogens* 9.11, e1003711.

Zhang, Ye, Larry H Rohde, and Honglu Wu (2009). "Involvement of nucleotide excision and mismatch repair mechanisms in double strand break repair". In: *Current genomics* 10.4, pp. 250–258.

Zheng, Christina L et al. (2014). "Transcription restores DNA repair to heterochromatin, determining regional mutation rates in cancer genomes". en. In: *Cell Rep.* 9.4, pp. 1228–1234.

Zheng, Huyong et al. (2003). "Nucleotide excision repair-and polymerase η-mediated error-prone removal of mitomycin C interstrand cross-links". In: *Molecular and cellular biology* 23.2, pp. 754–761.

Zhivagui, Maria et al. (2019). "Experimental and pan-cancer genome analyses reveal widespread contribution of acrylamide exposure to carcinogenesis in humans". In: *Genome research* 29.4, pp. 521–531.

Zou, Xueqing et al. (2018). "Validating the concept of mutational signatures with isogenic cell models". In: *Nature communications* 9.1, p. 1744.

Appendix A

List of DNA repair associated genes used in *C. elegans* mutagenesis screen

Gene name	**Pathway**	**Function of the corresponding protein**	**Human ortholog**
agt-1	Alkylguanine DNA-alkyltransferase	*MGMT*	
agt-2	Alkylguanine DNA-alkyltransferase, predicted to have transferase activity	-	
brc-1	DSBR	Ubiquitin protein ligase, required for HRR	*BRCA1*
brd-1	DSBR	Ubiquitin protein ligase, required for HRR	*BARD1*
bub-3	Spindle assembly checkpoint	Mitotic checkpoint protein	*BUB3*
ced-3	Damage checkpoint	Caspase, a cysteine-aspartate protease, required for execution of apoptosis	*CASP3*
ced-4	Damage checkpoint	Required for initiation of programmed cell death	*ARAF1*

cep-1	Damage checkpoint	P-53-like protein, promotes DNA-damage induced germ cell apoptosis, required for normal meiotic segregation in the germ line	*p53*
cku-80	DSBR	Predicted to contribute to double-stranded DNA binding activity; is involved in NHEJR	*KU80*
csb-1	NER	Functions in the transcription coupled NER, essential for overcoming deleterious effects of UV radiation	*CSB/ERCC6*
dna-2	Helicase	DNA replication helicase	*DNA2*
dog-1	Helicase, ICLR	Required for interstrand cross-link repair and for maintenance of poly-G tracts of germline and somatic DNA by resolving the secondary structure that can occur in G-rich DNA during lagging-strand DNA synthesis	*BRIP-1*
exo-1	BER, MMR	Exonuclease, implicated in MMR and DNA recombination	*EXO1*
exo-3	BER	Exhibits DNA-(apurinic or apyrimidinic site) endonuclease activity and phosphoric diester hydrolase activity	*APEX1*
fan-1	ICLR	Fanconi anemia-associated nuclease	*FAN1*
fcd-2	NER, ICLR	Predicted to have DNA polymerase binding activity, involved in NER and regulation of DNA-dependent DNA replication	*FANCD2*
fnci-1	ICLR	Predicted to have DNA polymerase binding activity	*FANCI*
fncm-1	ICLR	Required for resistance to DNA interstrand crosslinking agents	*FANCM*
gen-1	DSBR	Exhibits crossover junction endodeoxyribonuclease activity, is involved in HRR	*FANCD*

helq-1	ICLR	Putative DNA helicase, may unwind regions of cross-linked DNA before their repair by other proteins or resolve Holliday junctions after homologous recombination	*HELQ*
him-6	Helicase	RecQ-like ATP-dependent DNA helicase, required for replication	*BLM*
lem-3	Damage checkpoint	Protein containing ankyrin repeats, a LEM (LAP2-emerin-MAN1) domain and a GIY-YIG nuclease domain, involved in mediating DNA damage response	*ANKLE1*
lig-4	DSBR	Predicted to have ATP binding activity, DNA binding activity, and DNA ligase (ATP) activity, is invlved in NHEJR	*LIG4*
mlh-1	MMR	MutL homolog 1	*MLH1*
mus-81	DSBR, ICLR	Exhibits enzyme binding activity, is predicted to contribute to crossover junction endodeoxyribonuclease activity	*MUS81*
ndx-4	BER	Exhibits 5-phosphoribosyl 1-pyrophosphate pyrophosphatase activity and bis(5'-nucleosyl)-tetraphosphatase (asymmetrical) activity	*NUDT2*
parp-1	BER	Exhibits NAD+ ADP-ribosyltransferase activity; is involved in protein poly-ADP-ribosylation	*PARP1*
parp-2	BER	Exhibits NAD+ ADP-ribosyltransferase activity, is involved in protein poly-ADP-ribosylation	*PARP2*
pms-2	MMR	Mismatch repair endonuclease	*PMS2*
pole-4	Replicative polymerase	Accessory subunit of DNA polymerase epsilon	*POLE4*
polh-1	TLS	Ortholog of human DNA-directed DNA polymerase eta	*POLH*
polk-1	TLS	Ortholog of human DNA-directed DNA polymerase kappa	*POLK*

polq-1	DSBR, TLS	Ortholog of human DNA-directed DNA polymerase theta, involved in microhomology-mediated end joining repair	*POLQ, POLN*
rad-51	DSBR	Ortholog of human RAD51 recombinase, involved in multiple processes including HRR	*RAD51*
rcq-5	Helicase	Predicted to have several functions, including DNA binding activity, RNA polymerase II complex binding activity, and nucleoside-triphosphatase activity	*REQ5*
rev-1	TLS	Ortholog of human REV1 - DNA-directed translesion synthesis polymerase	*REV1*
rev-3	TLS	Ortholog of human REV3L - DNA-directed REV3-like DNA polymerase zeta, catalytic subunit	*REV3L*
rfs-1	DSBR	Rad-51 like protein, exhibits ATPase binding activity, involved in HR	*RAD51D*
rip-1	DSBR	RFS-1 interacting protein, involved in HRR	-
rtel-1	Helicase	ortholog of human RTEL1 (regulator of telomere elongation helicase 1), is predicted to have ATP binding activity, ATP-dependent DNA helicase activity, and DNA polymerase binding activity	*RTEL1*
san-1	Spindle assembly checkpoint	Ortholog of human BUB1 mitotic checkpoint serine/threonine kinase B, is predicted to have protein kinase activity	*BUB1B*
slx-1	DSBR, ICLR	Exhibits enzyme binding activity; contributes to 3'-flap endonuclease activity, 5'-flap endonuclease activity, and crossover junction endodeoxyribonuclease activity	*SLX1*

smc-5	DSBR	Ortholog of human SMC5 (structural maintenance of chromosomes 5) protein, is predicted to have ATP binding activity	*SMC5*
smc-6	DSBR	Ortholog of human SMC6 (structural maintenance of chromosomes 6) protein, is predicted to have ATP binding activity	*SMC6*
smg-1	Damage checkpoint	Ortholog of human SMG1 nonsense mediated mRNA decay associated PI3K related kinase	*SMG1*
tdp-1	BER	Ortholog of human TARDBP (TAR DNA binding protein), exhibits single-stranded RNA binding activity	*TDP1*
ung-1	BER	Exhibits uracil DNA N-glycosylase activity	*UNG*
wrn-1	Helicase	Exhibits ATP-dependent 3'-5' DNA helicase activity, involved in DNA metabolic process, determination of adult lifespan, and response to ionising radiation	*WRN*
xpa-1	NER	Ortholog of human XPA protein, DNA damage recognition and repair factor	*XPA*
xpc-1	NER	Ortholog of human XPC complex subunit, DNA damage recognition and repair factor	*XPC*
xpf-1	NER	Ortholog of the endonuclease catalytic subunit of human ERCC excision repair protein 4	*XPF/ERCC4*
xpg-1	NER	Ortholog of human ERCC excision repair protein 5	*XPG/ERCC5*

Table A.1: List of DNA repair genes knocked out in the screen.

Comments. *bub-3* and *polh-1* knockouts were made for two alleles each. All of the *brc-1* mutants also had a knockout of *brd-1*, and can be considered double knockouts brc-1, brd-1.

Appendix B

Mutational signatures of DNA repair deficiencies and genotype-genotoxin interactions in *C. elegans*

Experimental signatures of DNA repair deficiencies for all genetic backgrounds are available among other Supplementary Materials of Volkova et al. 2019 at the project's GitHub page: http://github.com/nvolkova/signature-interactions/Supplementary_tables (Supplementary Note Figure 1).

List of estimated interaction effects for 113/196 combinations of genetic backgrounds and genotoxins that showed a fold-change in mutation rate (total mutation rate, base substitution, indel or structural variant rates) different from 1 (FDR 10% in each category) is available among other Supplementary Materials of Volkova et al. 2019 at the project's GitHub page: http://github.com/nvolkova/signature-interactions/Supplementary_tables (Supplementary Note Figure 2).

Appendix C

Selection in DNA repair related genes and pathways across cancers

The list of DNA repair genes used in the study is available among other Supplementary Materials of Volkova et al. 2019 at the project's GitHub page: http://github.com/nvolkova/signature-interactions/Supplementary_tables (Supplementary Table 5).

The tables with dN/dS values per gene across cancer types and per DNA repair pathway per cancer type are available among other Supplementary Materials of Volkova et al. 2019 at the project's GitHub page: http://github.com/nvolkova/signature-interactions/Supplementary_tables (Supplementary Table 5).

Appendix D

Publications

List of publications during PhD studies

Bettina Meier, **Nadezda V Volkova**, Ye Hong, Pieta Schonfield, Peter J Campbell, Moritz Gerstung, Anton Gartner. (2018). Mutational signatures of DNA mismatch repair deficiency in *C. elegans* and human cancers. *Genome Research* **28**, pp. 666-675.

Santiago González, **Nadezda Volkova**, Philip Beer, Moritz Gerstung. (2018). Immuno-oncology from the perspective of somatic evolution. *Seminars in cancer biology* **52**, pp. 75-85.

List of submitted manuscripts

Nadezda V Volkova, Bettina Meier, Víctor González-Huici, Simone Bertolini, Santiago Gonzalez, Federico Abascal, Iñigo Martincorena, Peter J Campbell, Anton Gartner, Moritz Gerstung. (2019). Mutational signatures are jointly shaped by DNA damage and repair. *bioRxiv*, doi: https://doi.org/10.1101/686295.

List of manuscripts in preparation

Meier, B., **Volkova, N.V.**, Hong, Y., Wang, B., Gonzalez-Huici, V., Bertonlini, S., Boulton, S., Campbell, P.J., Gerstung, M. and Gartner, A. Systematic analysis of mutational spectra associated with DNA repair deficiency in *C. elegans* mutation accumulation lines.

Bettina Meier, **Nadezda V Volkova**, Victor Gonzalez-Huici, Peter J Campbell, Moritz Gerstung, Anton Gartner. Massive *C. elegans* whole-genome sequencing profiling to accumulate mutational signatures of alkylating agents and DNA repair deficiency.

Bettina Meier, **Nadezda V Volkova**, Simone Bertolini, Victor Gonzalez-Huici, Peter J Campbell, Moritz Gerstung, Anton Gartner. Systematic characterization of genome-wide mutagenesis conferred by ionizing radiation in wild-type and DNA repair defective *C. elegans*.